Automatic Assessment of Social Skills

DOCTORAL THESIS

for the Degree of a
Doctor of Informatics

AT THE FACULTY OF ECONOMICS,
BUSINESS ADMINISTRATION AND
INFORMATION TECHNOLOGY
OF THE
UNIVERSITY OF ZURICH

by

ROBERT STOYAN
from Germany

Accepted on the recommendation of
PROF. DR. HELMUT SCHAUER
PROF. DR. JULIUS KUHL
PROF. DR. MICHAEL HESS
2012

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich herewith permits the publication of the aforementioned dissertation without expressing any opinion on the views contained therein.
Zurich, April 4, 2012

The Vice Dean of the Academic Program in Informatics: Prof. Dr. Harald C. Gall

Acknowledgements

I would like to sincerely thank Prof. Dr. Helmut Schauer for his guidance, understanding and inspiration while working on my PhD thesis at Zurich University. I am not sure how many PhD students are given the independence to develop their own research directions.

My thanks are equally due to Dr. Tsuyoshi Ito, my closest colleague at Zurich University, with whom it was a great experience to work with on the PM Game. Although our contributions to the PM Game were clearly separate, being the basis of two diferent PhD projects, our daily discussions were inspiring and I could always rely on Tsuy. I greatly appreciate the high quality work of Boris Otonicar, Michael Brändli, Dr. Julian Ebert, Philipp Fonberg and Sudeeptha Grama Visweswara. Boris served as an assessment centre observer, and Michael both as role-play actor and assessment centre organiser. In this assessment centre, Julian's role-play scripts were used. Sudeeptha and Boris validated behavioural markers. Philipp designed the layout of the PM Game. Now, at the time of writing this thesis document, I am glad to continue the work on the PM Game together with Ioana Ilea at Geneva University.

I am also indebted to Prof. Dr. Julius Kuhl and Prof. Dr. Michael Hess for refereeing this thesis, and equally to Dr. Arlena Jung, Prof. Dr. Gisbert Stoyan, Dr. Beatrice Hasler, Dr. Markus Quirin and Dr. Nicolas Szilas for many helpful feedbacks on the entire or parts of this thesis.

Prof. Dr. Martin Glinz financially supported this project during a critical phase. One of his lectures inspired the contents of the game story 'Mary and Garry' on eliciting customer requirements. Many thanks are also due to Prof. Dr. Gerhard Schwabe for popularising Hevner's design science approach at the Institute for Informatics. The insights from Prof. Hevner's talk on design science formed the structure of this thesis document. Prof. Dr. Michael Hess, Prof. Dr. Julius Kuhl, Dr. Julian Ebert, Dr. Beatrice Hasler and Prof. Dr. Kleinmann gave helpful advice on concepts of computational linguistics, the PSI model, assessment of social skills, virtual assessment centres and assessment centres, respectively.

I owe many thanks to Dr. Katrin Häsler for her invaluable advice and support regarding formal requirements.

I especially wish to express my thanks to my sponsors, who provided substantial financial means to support my research: Gebert Rüf Foundation, Hasler Foundation, the Swiss National Science Foundation and McKinsey. Without their generous, and in several cases, repeated support, this research would never have been possible.

I would like to thank the 40 students who participated in my research in various forms from theses to study projects during my years at Zurich University. Many thanks are equally due to the Institute of Informatics who made it possible for me to supervise such a large number of students.

Last but not least, I wish to express my sincere thanks to my family for their support during this work: Hajna, Gisbert and Helmar Stoyan.

Contents

Index of abbreviations

NLI	natural-language interface
NPC	nonplayer character
r	correlation coefficient
SCT	sentence-completion test
SJT	situational judgement test

Index of figures

Index of tables

Zusammenfassung

Das Thema dieser Dissertation ist ein Computerspiel, welches dem Nutzer die Interaktion mit simulierten menschlichen Agenten ermöglicht. Die Dissertation belegt empirisch, dass dieses Spiel bestimmte soziale Skills beurteilen kann. Dabei werden lediglich technische Mittel verwendet. Bislang war valide Beurteilung von sozialen Fähigkeiten durch Computer nicht möglich; menschliche Beurteiler waren erforderlich.

Die Dissertation beschreibt die technischen Methoden, die für Kommunikation, Simulation, Verhaltensbeobachtung und Fähigkeitsbeurteilung verwendet werden, sowie auch Anwendungsszenarien. Eine Hypothese wird aufgestellt und begründet, dass eine bestimmte Form der menübasierten Kommunikation das zentrale Element des Spiels ist, welches Beurteilung von sozialen Fähigkeiten ermöglicht. Das Spiel wird nur durch die Maus bedient. Der Benutzer interagiert mit einem Menü und einem Satzergänzungsmechanismus. Das Ziel dieser Lösung ist es, dem Nutzer Freiheiten im Ausdruck zu bieten und gleichzeitig für den Computer volle Sicherheit der Bedeutung zu gewährleisten. In diesem Sinne vereint das vorgestellte menübasierte Satzergänzungsinterface einen Vorteil von freier Sprache (Ausdrucksfreiheit) mit einem Vorteil von Multiple-Choice-Interfaces (Sicherheit der Semantik). Satzergänzungstests wurden in der Computerlinguistik verwendet, jedoch bislang nicht für die Beurteilung von sozialen Fähigkeiten genutzt. Bisherige nicht erfolgreiche computerbasierte Versuche soziale Skills zu beurteilen, verwendeten entweder freie Sprache oder Multiple-Choice.

Die Validität der computerbasierten Assessmentmethode wird durch eine Vergeichsstudie belegt: Die verbale Leistung soziale Situationen zu meistern wurde für 47 Studienteilnehmer sowohl vom Computerspiel als auch von menschlichen Beurteilern in einem etablierten Beurteilungsverfahren bewertet. Für diesen Vergleich wurden drei industrie-relevante soziale Fähigkeiten gewählt:

- Ein Feedback-Gespräch mit einem Angestellten führen
- In einem Erstgespräch einen emotionalen Kontakt mit einem Kunden knüpfen
- Anforderungen eines Kunden erfragen

Das Computerurteil korrellierte signifikant mit dem menschlichen Urteil. Ein Benchmarking-Vergleich zeigte, dass zwei der drei Korrelationen den Korrelationen zwischen etablierten Assessmentmethoden entsprechen (Parallelversionen von in der Forschung verwendeten Assessmentzentren). Zusätzlich wurde eine theoretische Analyse durchgeführt mit dem Ergebnis, dass computerbasierte Beurteilung mit der vorgestellten Methode möglich ist. Hierfür wurden Daten aus früheren Assessmentverfahren anderer Forscher verwendet. Schliesslich stellt die Dissertation Argumente bereit, dass die computerbasierte Methode auf weitere soziale Fähigkeiten erweitert werden kann.

Abstract

The subject of this dissertation is a computer game that simulates humans the user can communicate with. This dissertation claims and provides empirical evidence that this game can assess certain social skills of the user solely by technical means. Until now, valid assessment of social skills by computers has not been possible; human assessors were needed.

The dissertation describes the technical methods used for communication, simulation, behaviour observation and skill scoring as well as application scenarios. A reasoned hypothesis is made that a specific form of menu-based communication is the central feature of the game that enables the assessment of social skills. Being steered solely with the mouse, the game provides the user with a menu and a sentence completion mechanism in order to communicate with the game. The goal of this solution is to provide the user with relative freedom of expression and, at the same time, ensure that the computer has full certainty of the semantics of the human communication. In this sense, the proposed menu-based sentence completion interface combines an advantage of free speech (freedom of expression) with an advantage of multiple-choice interfaces (certainty of semantics). Although sentence completion tests have been investigated in computational linguistics, they have not been applied so far to assess social skills. Previous unsuccessful attempts at computer-based social skill assessment used either free speech or multiple-choice interfaces.

Evidence for the validity of the computer-based assessment method is provided by a comparative study: the verbal performance in mastering social situations was assessed for 47 study participants both by the computer game as well as by established means of assessment using human assessors. For this comparison, three industry-relevant social skills in communicating with subordinates and customers were selected:

- Conducting a feedback talk with an employee
- Establishing an emotional bond with a customer in a first customer meeting
- Eliciting customer requirements

The computer assessments correlated significantly with the human assessments. A benchmark-comparison was carried out showing that two of the three correlations are of the same size as best correlations between established means of assessment (parallel versions of research-quality assessment centres).

Additionally, a theoretical analysis was carried out using independent data from other researchers on preexisting assessment methods. It confirms the feasibility of computer-based social skills assessment by the method proposed in this thesis. Finally, argumentative evidence is provided that the computer-based method can be extended to further social skills.

Overview of the research fields involved

This dissertation involves the following fields of research:

- Computer science: computer-based skill assessment
  (in particular: virtual assessment centre)
- Computer science: serious computer games
  (in particular: human behaviour simulation and digital storytelling)
- Computational linguistics
  (in particular: menu-based communication interfaces and sentence completion tests)
- Psychology
  (in particular: test validation and assessment centre psychology)

1 Introduction

Computer-based assessment has a long history of research. In the assessment of job applicants, attention was mainly given to multiple-choice tests and to computer games to test problem-solving capabilities (Section 3.1.1). Research on assessment for the purpose of education has focused on computerising exams, mostly using multiple-choice. A rich variety of computer tests have been developed assessing IQ, verbal, practical and social intelligence, personality, motives, task prioritisation, language pronunciation and knowledge, problem-solving capabilities in many job areas, and knowledge on almost every imaginable subject. There has been, however, one skill area for which attempts to develop and validate a computer-based assessment method have been unsuccessful: social skills (Section 3.1.1). Probably the first attempt that may fairly be considered 'testautomatisation' was the SI-Test by Moss et al. (1927). Its purpose was to test the 'ability to get along with people', which is a claim to test social skills. It was a multiple-choice test on judgment in social situations, name and face memory, observation of human behaviour, recognition of mental states and sense of humour. Looking at the test contents, it seems plausible that it could be capable of assessing aspects of social intelligence such as name and face memory, and possibly also a spectrum of social skills including behavioural aspects. However, it failed to prove validity (Section 3.1.1). Why has there been little progress since, in spite of the development of computers? Plausible explanations include the fact that computers are still limited in their ability to understand human communication or the fact that cognitive models and artificial intelligence are still far from unravelling the complexity and performance of the human mind. This thesis does not attempt to solve these fundamental challenges. In fact, it takes an approach in the opposite direction. The method proposed in this thesis can be described as a technology developed to achieve social skill assessment by circumventing natural-language understanding (Section 4.5) and artificial intelligence (Section 4.8.2). It is a means of doing what is necessary for social skills assessment but nothing more: providing the stimuli to make human test takers reveal their social behaviour in a computer observable fashion. Photorealistic graphics, 3D, real time, movement, and other such elements are avoided. The entire system is kept simple and authorable (Section 4.8). It is reasonable to assume that it can be extended (Section 5.6) to social skills not tested in this thesis. Compared to high-fidelity simulation (e.g. immersive virtual computer worlds or role-plays with humans) and Motowidlo's (1990) concept of low-fidelity simulation (social judgment tests using multiple-choice), the solution proposed in this thesis could be called a 'medium fidelity simulation' of social situations and conversations therein.

1.1 Definition of the research problem

The research task carried out in this dissertation consists of developing a computer-based method for the assessment of the verbal aspects of social skills, validating it for selected industry-relevant social skills and drawing conclusions regarding the method. Thus, the thesis addresses a measurement problem by solely technical means, which has been addressed so far by assessment psychology using trained human workforce.

To avoid raising inflated expectations, it is pointed out clearly that the method is not claimed to assess all social skills. It is empirically validated to assess the verbal aspects of three social skills, and argued to be extendable to several other social skills. Thus this thesis is a proof-of-concept study. As such, it is not about optimising and perfection.

1.2 The object of measurement/assessment

Having the research problem stated as being a measurement problem, the question arises what the exact object of measurement is. In this section, it will be examined what skills and social skills are and it will be defined what will serve as the criterion for sufficient measurement accuracy.

Skills generally refer to the ‘ability to perform’ (Spitzberg & Cupach, 1989). Thus, for example, assessment of typical social behaviour and assessment of motivation do not count as assessment of social skills. Are skills measurable? Having defined 'skill' as the 'ability to perform', a distinction is made between the performance, which has an effect in the world, thus, is potentially measurable, and the skill or ability, which is a concept (a 'construct' in the terms of psychology) that has in itself no effect in the world. Thus, when 'assessing skills', performance is measured and hoped to give insights about the skill that enabled that performance. For simplicity, in the rest of the thesis, generally the term 'assessment of (social) skills' will be used, though what the method actually does is measurement of the performance shown in a specific test session.

While the definition of skills is well accepted, there is no generally agreed definition of social skills. Hasler (2009) lists the following definitions:

Feldman, Philippot, & Custrini (1991, p. 331)	Social competence is ‘a hypothetical construct relating to evaluative judgments of the adequacy of a person's social performance. . . . We assume that specific social skills, overt or cognitive behaviors that an individual performs, underlie social competence’.
Ford (1982, p. 323–324)	‘Social competence is the attainment of relevant social goals in specified social contexts, using appropriate means and resulting in positive developmental outcomes’.
Hargie (2006, p. 13)	‘Social skill involves a process in which the individual implements a set of goal-directed, interrelated situationally appropriate social behaviours, which are learned and controlled’.
Oppenheimer (1989, p. 45)	Social competence refers to ‘the ability to engage effectively in complex interpersonal interaction and to use and understand people effectively’.
Rose-Krasnor (1997, p. 123)	‘Social competence is defined as effectiveness in social interaction’ and ‘as an organizing construct, with transactional, context-dependent, performance-oriented, and goal-specific characteristics’.
Segrin (1998, p. 229)	‘Social skills refer to the skills and abilities that allow people to interact appropriately and effectively with others. Social skills are, to a greater or lesser extent, manifested and exercised behaviorally whenever two or more people interact with each other’.
Spitzberg & Cupach (1989, p. 7)	‘Competence is manifested in effective and/or appropriate behavior’.

Table 1: Different definitions of social skills

A concept that has no generally accepted definition is hard to measure. Hence, for the purposes of this thesis, as a conclusion from the majority of the above definitions, the following will be assumed:

Social skills manifest in behaviour and they refer to interaction between persons. That is, the ability to form a judgment is not a social skill, even if it is judgment of social situations or behaviours therein. Remembering names, faces and relationship details or inferring about one's own or others' mental states and intentions, are also not social skills, though obviously useful for successful social behaviour. These exclusions will be important when investigating if previous research has reached the goal of assessing social skills, see Section 3.1.1.

Apart from the above decision to focus on behaviour in interaction with other persons, this thesis will not require any decision regarding what social skills are. Instead, the following criterion will be chosen to determine if the goal of assessing social skills by technical means is reached.

The technical assessment method proposed in this thesis will be regarded accurate if it produces similar results to the most established and best validated measure for social skills that uses human workforce, that is, assessment centres. Similarity will be measured in the same way as similarity between outcomes of established methods is measured, that is, by correlation.

Last but not least, as stated in the definition of the research problem in Section 1.1, the focus of the dissertation are the 'verbal aspects of social skills'. This means that only performance is measured that manifests in words. This applies both to the technical method to assess social skills as well as to the human assessment to compare with. All other means of expression such as voice, mimics, gesticulation and posture are ignored.

1.3 Structure of the dissertation

The structure of this thesis follows the design science paradigm as described by Hevner (2004, p. 76). 'The design-science paradigm has its roots in engineering and the sciences of the artificial (Simon 1996). It is fundamentally a problem-solving paradigm. It seeks to create innovations…' 'Design-science research addresses important unsolved problems in unique or innovative ways or solved problems in more effective or efficient ways. The key differentiator between routine design and design research is the clear identification of a contribution to the archival knowledge base of foundations and methodologies' (Hevner, 2004, p. 81). In short, a problem has to be defined, its importance needs to be demonstrated, an innovation has to be presented and finally evaluated if there is an addition to the knowledge base. The sequence of the main chapters of the thesis follows this approach:

1	Introduction (and definition of the research problem)
2	Relevance (of the research problem)
3	State of the art (pertaining to research that attempted to resolve the same problem)
4	Design (of the method/software for solving the problem)
5	Evaluation (of the proposed method)
6	Summary and conclusion (what the contribution to the knowledge base is)

Table 2: The main chapters, following design science concepts

2 Relevance

When proposing a new method, according to the design science approach (Section 1.3), its practical relevance has to be demonstrated. This is an invitation for the researcher of informatics to not only think within one's own field of research, but to visit the subject fields for which one is developing new computer-based methods and orient the technical development towards real needs.

2.1 Relevance of social skills

Researchers have identified social skills as one of the most important personal qualities needed for success on the job (Ferris, Perrewé & Douglas, 2002; Riggio, 1986; Riggio, Riggio, Salinas, & Cole, 2003; Witt & Ferris, 2003). Recent studies have suggested that graduates' greatest deficit is in social and interpersonal skills rather than in technological know-how (Taylor, 2006). In today's modern teamwork-oriented work processes, scientific and technical knowledge can only be applied effectively if combined with interpersonal skills (German Federal Ministry of Education and Research, 1999). These statements indicate the following practical needs:

Need	Use Cases
Employers need to select graduates and other applicants with regard to their social skills	1
Teaching/learning of social skills should be improved	2, 3, 4

Table 3: Practical needs and use cases for assessment of social skills

These needs will be analysed in more detail in the following sections. For each need, use cases for social skills assessment will be proposed.

2.2 Use Case 1: Computer-based job assessment

2.2.1 The practical problem

Assessment is carried out by organisations as a means of measuring the potential and actual performance of their current (post-hire assessment) and potential future employees (pre-hire assessment). Measurement is important because it enables organisations to act both tactically and strategically to increase their effectiveness (Bartram, 2004). Especially with regard to social skills, these assessments require the work of human assessors (Section 3.1.3), which results in substantial cost. For example, assessment centres cost 500-2500 Euros per person per day in some European countries (Hogan & Zenke, 1986; Gerpott, 1990; Hoffmann & Thornton, 1997). As assessment of skills is expensive, most applicants are currently turned down based on their CV, in spite of the fact that biographical data explains only 7.8% of the variance of job performance (weighted average of four meta-analyses; Bobko et al., 1999). Thus, companies lose capable applicants and applicants lose job opportunities for which they are qualified. As long as skill testing remains expensive, this situation will persist.

In short, the problem is essentially a cost issue forcing employers to turn down most applicants based on data that does not provide adequate means for estimating actual skills.

From the perspective of companies and applicants, the costs mentioned above are multiplied as the same applicant is assessed several times by the various companies he or she applies to.

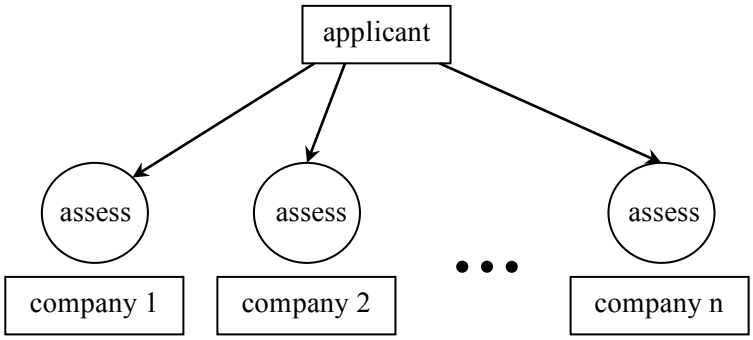


Figure 1: Repeated applicant assessment

The solution is simple in principle:

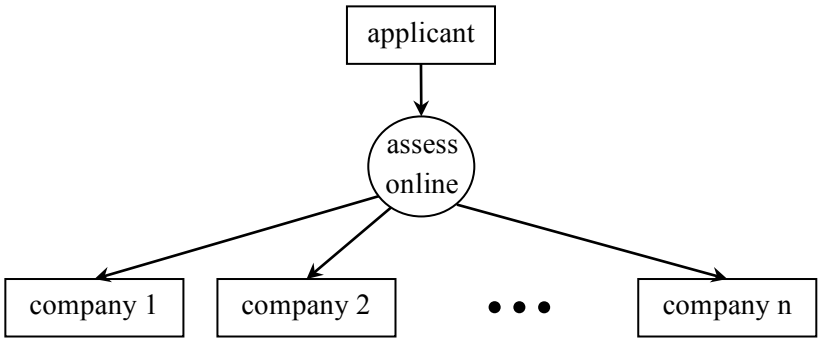


Figure 2: Central online assessment

Viewing the above diagram, the following questions arise: would a single assessment decide the fate of a person? Who controls the central testing entity? How does one prevent cheating, for example, by letting somebody else sit in front of the computer and take the tests? Fortunately, these questions have been resolved with the use of online testing systems (see below) that are already in operation. An issue that has not been resolved, however, is that the online tests lack social skills assessment. This constitutes the Use Case 1 for the social skills assessment method proposed in this thesis.

2.2.2 Examples of online testing services

Large-scale internet-based skill testing systems are in operation in several countries, but the largest one is in India, catering to foreign companies with regard to their need for software developers. India is the classic offshore IT development destination for many large international companies. As Wiener (2006) points out, for large software projects, ‘global delivery’ is regarded as the most successful offshore development method. This means that western companies do not outsource their software development directly to India, but to other companies that in turn outsource the bulk of the development (typically the detailed design and the coding) to their Indian branch. Filling these Indian branches with personnel creates a demand for an assessment of the applicants, which has called for the development of assessment services. The largest Indian company currently providing such assessment services is MeritTrac Ltd, www.merittrac.com. They operate a test centre network, selecting over 40,000 Indian IT professionals per month (MeritTrac, 2007). The main body of their tests are validated multiple-choice tests. The quality of English pronunciation is tested by voice analysis. English language skills are one of the major reasons for foreign companies to outsource software development to India (Wiener, 2006). The possibility of cheating by letting another person take the test is prevented by the operation of test rooms in every major Indian city.

Another innovative example is www.karrierejagd.de (Cyquest Ltd. in Germany), which operates an online recruiting platform (Karrierejagd, 2003). They claim to have over 40 company partners and to have had more than 110,000 users since 2000. Their largest user group are IT graduates. This platform asks applicants questions related to their CV in a playful manner. In addition, multiple-choice tests are provided, if requested by a company for the selection of their applicants.

In the case of the German example as well as several other European test services, a simple method is used for dealing with the risk that another person could take the test: the online test only functions as a pre-screening. Candidates who pass the online test are invited to a personal interview and further tests. In this procedure, having the pre-screening test completed by another person not only provides little advantage, but also involves an additional risk for the applicant: if the other person provided strongly differing statements e.g. in a personality test, the application could be questioned. The advantage of this selection procedure is that the test room infrastructure where IDs are controlled is not necessary. The pre-screening tests are taken at home.

2.2.3 The necessity for selection

In the previous section, examples of online testing services were presented, mentioning the large number of applicants assessed. This is one of the factors driving the need for online testing services. A further key factor is the high selection rates. While high selection rates can be observed worldwide, the reasons for their occurrence differ, depending on the country and job:

In India and China, the employability of graduates is low, causing substantial need for selection (Gerefi et al., 2008). A survey by the McKinsey Global Institute (2005) measured the employability of engineers in global labour markets. McKinsey surveyed human resource professionals from 83 companies operating all over the world, and asked them the following question: ‘Of 100 [engineering] graduates with the correct degree, how many could you employ if you had demand for all ?’ Respondents stated that 80.7 per cent of U.S. engineers were employable, while only 10 per cent of Chinese engineers and 25 per cent of Indian engineers were similarly employable.

In western countries, jobs with a highly competitive selection process can be found for example in the management field. Selection rates for management positions in Germany are typically under 1% (StepStone, 2004). In the example of the company UniLever, 5000 applicants apply annually for project management positions (Kopping, Diercks & Kupka, 2007). However, around 40 are needed per year. Another example for competitive selection is trainee positions. Here, selection rates are equally under 1% (StepStone, 2004). Generally, formal qualifications can easily be checked; however, of interest are actual skills. In the case of managers, the point of interest for this thesis is that their job is, in large part, communication. Hence, social skills are needed. In the case of trainees, a particular point of interest is that they usually do not have a long history of work experience. Thus, the difference among trainees can only be identified by assessing their skills.

A further interesting factor driving the need for online selection is in fact the Internet itself: applications by email have caused the total number of applications to rise sharply (StepStone, 2004).

Currently, companies are in the process of discovering online assessment. While in 2008, only 7.6% of the 1,000 biggest German companies were reported to use online assessment, a further 7.6% reported that they planned to use it (Laumer et al., 2009).

Discussion and conclusion

The above data shows the need for online assessment in general as well as specifically for social skills. Thus, there is justification for Use Case 1. However, is this use case realistic? There is a fundamental difference between social skills and other skills when it comes to practically implementing this use case: a company can easily specify their need for a certain number of software developers, for example, with the requisite software development skills. However, how can they specify what they need in terms of social skills? It is common knowledge that specifications of required social skills in job advertisements are vague. For example, 'good team player' can mean just about anything depending on the leadership style, the company culture, the other members of the team, etc. Furthermore, it is unrealistic to expect companies to explicate their needs. Is there a company that would state: 'We need a software developer who can interact well with a conservative commanding boss in an overregulated company'? However this social skill, as negative as it sounds, is an existing need. Furthermore, the manager in question can assess this skill in a personal interview. If an assessment centre were used, the HR specialist of the company would have incorporated the company culture into the design of the assessment centre exercises. Consequently, the need to articulate the above negative statement would not arise. Computer-based assessment for social skills, being a technical innovation, differs in these regards from traditional assessment centres. It would definitively not be created by the employer as an assessment centre would. The employer would have to provide the requirements to the online testing service. It may be another company or it may be a piece of software that is rented or bought. In any case, there is need for explicating the requirements. Thus, summarizing this paragraph, computer-based assessment raises issues that were either not present at all or not present to that extent when using traditional means of social skills assessment. Proposing a method for assessment is only the beginning of a long journey.

Finally, assessment procedures are not just an assessment of the applicant. They also convey an image of the company, which may attract or scare away applicants. Traditional means of social skills assessment leave room for dealing with this risk, for example, by inviting the employee to a presentable company building. There are, however, possible ways of dealing with these issues when using online assessment of social skills. The values represented by the company could, for example, be incorporated in the behaviour of the non-player characters that the applicant meets in a game such as the one presented in this thesis.

The above discussion shows that several important issues have to be dealt with in order to implement Use Case 1 in an industrial setting.

2.3 Use Case 2-4: Using computer-based assessment to improve social skills education

This section examines the issues of social skills education in companies and educational institutions and presents use cases for computer-based assessment to improve the situation.

2.3.1 Issues in social skills education

Social skills education takes place at many levels. Companies often offer leadership and customer communication trainings to their employees. Classes in schools and universities use teamwork activities. However, there is one substantial difference compared to other areas of learning: social skills are rarely assessed. In school, marks are rarely given for social skills. Project management certificates are awarded by assessing project planning, budgeting, milestone tracking, and similar skills; however, there is little or no assessment of the ability to lead a team. Recent initiatives in many countries have

promoted social skills education in kindergarten and school. However, the lack of assessment decreases the importance of social skills education. Unlike a bad mark in mathematics, poor social skills are of little hindrance when progressing to the next class or the next educational level. This is ironic because most people end up using just simple mathematical calculations, but everyone needs social skills in their work and private lives. In companies, the main area of post-hire assessment where instrumentation has been researched in some detail is that of management and leadership assessment. In most other areas, there has been a tendency for organisations to be far less concerned about the quality and validity of the tools they use for post-hire assessment (e.g. for 360-degree feedback) than they are about those they use for selection (Bartram, 2004).

2.3.2 Use Cases 2-4

In view of the issues presented above, improvements could be made both by supporting traditional means of education as well as by creating computer-based means of education:

Use Case 2: Provision of objective and unbiased assessment of skills to determine what particular social skills are well developed or lacking in an individual. Thus, conventional training can focus on the most important deficits.

Use Case 3: Evaluation of the difference in social skills before and after a course. Among others, this can be used to provide grades for the participant or to determine the effectiveness of the educational intervention.

Use Case 4: Assessment as part of e-coaching. Currently, the author of this thesis is conducting a research project to utilize the means of social skills assessment described in this thesis to develop an e-learning solution for social skills. The idea is to assess social skills in an individual and provide the assessment results in an educationally effective fashion with detailed feedback, hints for improvement, and pointers to theory. Overall, the role of the computer would be similar to that of a human coach. Individual coaching is regarded as being most effective for learning (Corbett, 2001). Ohlsson (1994) and Corbett (2001) have shown that e-coaching can achieve learning effects that equal or supersede those of individual human coaching. The learning objectives were problem solving skills and programming skills.

Similar to Use Case 1, the above use cases will have their list of practical challenges when implemented in the future. The provision of the method and the game presented in this thesis is only the first step. Further challenges to be resolved include the following:

- Use Case 2 and 3: aggregation of scores for individual skills to higher-order skills
- Use Case 3: retest issues (if people take the same test twice, they may improve. Thus a learning effect may be observed, even if the educational intervention between the two tests contributed nothing to it)
- Use Case 4: transfer problems (will the learning effect achieved in the game transfer into real world behaviour?)

2.4 Discussion and conclusion

Four use cases have been presented for computer-based assessment of social skills. While these address different areas of application, as a common feature they all address large numbers of users. Large numbers of users are an important condition for successful e-learning applications (Breitner, 2006), and the same can be reasonably argued to be a prerequisite for successful computer-based as-

sessment (e-assessment) because the development of software and content is typically expensive. For the same reason, if learning content (or equally assessment content) needs to be updated or replaced frequently, that makes them less suitable for e-learning (e.g. Breitner, 2006). The issue is readily illustrated by computer and software technology as a learning field: not only the technologies themselves, but even basic methodical knowledge changes every couple of years. An impressive example for this is how technological developments in cloud-storage have made nearly all existing database methodology redundant (cf. Florescu & Kossmann, 2008). In contrast, social behaviours are stable over long periods of time. Thus, a test battery, once developed for one of the presented use cases, may be expected to have a long use period.

In conclusion, social skills are, both with regard to user numbers and the stability of content, a promising field for computer-based assessment.

However, there are also challenges to consider. Some of these have already been presented while discussing the different use cases above. Further, challenges common to all the use cases include the following:

- Substantial amounts of content are needed in all cases. An obvious reason is that in order to cover large skill areas such as leadership skills, many sub-skills such as delegating or giving feedback to employees, moderating meetings, conducting conversations to motivate people or to warn or even fire them, and many more must be considered.
- Several tests must be available for the same content: In Use Case 3 there needs to be pre- and post-tests for the same learning content; in Use Case 1 and 2 there is a need to refresh test contents on a regular basis in order to avoid familiarity. In Use Case 4, there could be learners wishing to reinforce a lecture by being confronted with different situations to learn the same skills.

The technical implications of the above are: good extensibility of the assessment method and the means to create parallel versions. Such means already exist for established forms of testing, compare for example, Irvine & Kyllonen (2002); these are often termed ‘test item generation methods’. However, in this thesis, we will use the general IT term ‘extensibility’, implying the addition of new skills and new stories alike. Extensibility of the method for social skills assessment proposed in this thesis will be addressed in Section 5.6. The possibility of creating parallel versions of stories will not be directly addressed in this thesis document. However, the validation study (Section 5.2) for the proposed computer-based assessment method has solved a similar, though more challenging task: to create computer tests that are parallel versions of traditional human-assessed role-plays.

3 State of the art

So far, there has been no computer based assessment method that has been validated to assess social skills. In order to substantiate this and in order to provide a broad overview of social skill assessment methods, the state of the art will be presented in the following parts:

- Section 3.1 State of the art of computer based social skill assessment
  - Section 3.1.1: Failed attempts to assess social skills
  - Section 3.1.2: Methods assessing similar constructs
- Section 3.2: State of the art of social skill assessment relying on human judgment

Additional information related to the state of the art

In Section 1.3, design science was set out as the paradigm for structuring this thesis. In addition to a state of the art as presented in this chapter, design science gives also emphasis to the practical aspects and looks at research as a means to add to the methodological knowledge base of the field. These are contained in the Chapters 2 Relevance and 5 Evaluation:

- In order to practically motivate the work contained in this thesis, Section 2.2 and 2.3 reviews online testing services currently in use and the practical problems of educational assessment.
- Section 4.2 presents the constituents of the assessment method used in this thesis. Building on this, in Section 5.1, prevalence of these methodical elements is searched among existing computer based methods independently of their purpose. This pre-existing research did not target social skill assessment but mostly education. The question is essentially if this thesis uses a known method for a new purpose or it contains new methodology.

3.1 State of the art of computer-based social skills assessment

An abundant literature on computer-based assessment in general provides interesting results, such as evidence that they can be designed to be fair and acceptable (e.g. Escudier, Newton, et al., 2011), just to mention one typical representative of current research topics. Further results have been produced over the last decades of research on social skills assessment using human judgment, as summarised in Sections 3.2.1 and 3.2.2. However, the question for computer-based social skill assessment is rather basic: has it produced any significant results?

3.1.1 Attempts to assess social skills

Attempts have been made to assess social skills using computers and paper-and-pencil multiple-choice tests, as will be outlined below.

Remark on multiple-choice tests: while multiple-choice tests are not always administered and scored by computers, all can be considered ‘automatic tests’, as the human work can always be done by computers. In fact, multiple-choice tests administered by computers and on paper have been shown to be equivalent: the increasing use of computerised testing has inspired research on the equivalence of computer and paper-and-pencil tests that has generally indicated that tests, especially un-timed self-report measures, need not be affected by this change of medium (Bartram, 1994; Donovan, Drasgow, & Probst, 2000; Neuman & Baydoun, 1998). However, high-speed ability tests (such as clerical checking tasks) may need renorming, as the test ergonomics can significantly affect the speed with which people can perform it. In sum, all multiple-choice tests are considered ‘relevant competitors’ of the assessment method proposed in this thesis.

Multiple-choice tests

Situational judgment tests (SJT) present a textual description of a situation and request a choice from predefined alternatives. Thus, they are multiple-choice tests for social situations. They have been tried in the automatic assessment of social skills in several studies. Perhaps the first attempt was the SI-Test by Moss et al. (1927), whose purpose was to test the 'ability to get along with people', which is a claim to assess social skills. It was a multiple-choice test on judgment in social situations, memory for names and faces, observation of human behaviour, recognition of mental states, and sense of humour. However, what was mostly assessed was verbal intelligence (Woodrow, 1939, cited in Probst, 1975; compare also Thorndike & Stein, 1937). Donovan (1998) presents a more typical recent example of an SJT where social skill assessment is explicitly stated as the purpose. The validity of its assessment of social skills has, however, not been established. The most successful SJTs have achieved correlations with measures related to social skills. For example, Motowidlo et al. (1990) achieved correlations between several overall SJT scores and supervisory ratings for 'interpersonal effectiveness' and 'communication effectiveness' (44** [$p < .01$] and .43 [$p < .01$]). While this constitutes evidence of the valid prediction of job performance, it does not validate the assessment of social skills, as supervisory ratings are well known to share a substantial variance with knowledge, practical intelligence, and intelligence: in knowledge tests, practical intelligence is measured by the ability to guess hidden assessment centre criteria, and intelligence test scores have been shown many times to correlate with job performance as measured by supervisory ratings (Kleinmann, 1997). Thus, one cannot determine from these results which skill was involved. Motowidlo et al. did not claim that their SJT would assess social skills and did not attempt to prove that it would (exclusion of the assessment of intelligence, etc.). Meanwhile, sound meta-analytic evidence confirms that SJT can be designed and validated to assess practical intelligence, general intelligence, personality traits and knowledge (Chan & Schmitt, 2005; McDaniel et al., 2007). There is no evidence of its validity for the assessment of social skills, however, in spite of decades of research.

This situation is best portrayed by a study of Funke & Schuler (1998) in which SJT and role-plays were developed to assess behaviour in the same six social situations. However, they correlated non-significantly by $r = .13$ for the SJT oral stimulus and $r = .17$ for the video stimulus. In comparison, correlations between the role-plays and situational interviews (developed for the same six situations) were .59***. Further details of the study are described in Section 5.5.1.

In view of the above evidence, it appears to be reasonable to conclude that it is not possible to assess social skills using SJTs. In other words, data on choices a subject makes from predefined alternatives in a social situation apparently do not contain information on the subject's social skills.

Repeated multiple choice or numerical interaction

Classical multiple-choice tests as cited above require the subjects to produce a single answer. There may be advantages to social skills assessments using repeated interaction.

O'Neil, Allred & Dennis (1994, 1997) reported evidence for the validity of a computer simulation assessment of negotiation skills. One of their titles appears to claim the development of a computer-based method of assessing social skills: 'Validation of a computer simulation for assessment of interpersonal skills'. Has a social skill been assessed through the assessment of negotiation skills? The fact that no verbal communication was involved in the assessment should not exclude the possibility of social skill assessment: social interaction does not happen only through words; the task was to choose

reasonable counteroffers; the human subject and the computer submitted numerical bids to counter each other). Furthermore, negotiation skill unquestionably fulfils the definition of social skills (see Section 1.2). This does not imply, though, that the numerical aspect that has been assessed is a social skill. As an analogy, the skill of parting goods in a social situation is a social skill, but this does not make the ability to compute a certain fraction of a sum a social skill as well. More generally, not only social skills but many other skills may contribute to the successful resolution of a social situation. Such skills include cognitive performance, language abilities, factual knowledge, and many others. This is why the validation of social skill assessment must analyse how much of the variance in the putative social skill scores stems from other, non-social-skill sources of variance. Such an analysis was not carried out, and there is thus no evidence of social skills assessment. By no means does this dispute the validity of their assessment method for negotiation skills, or its practical applicability, or the value of the results. It is simply to conclude that the claim of social skills assessment that may be inferred from one of the titles has not been substantiated. While some of the people the author talked to understood their research as having assessed social skills, reading the original papers makes it apparent that this claim was likely not intended by the authors: the articles do not contain such a claim. The title might have been meant to say 'we assess negotiation skills and negotiation skills are social skills (to some extent)' and not, as it literally reads, 'we have assessed social skills'.

Verbal interaction

Paschall et al. (2005) have developed a video game in which the user can interact with a non-player character through speech. The purpose of the game is to assess the social skills of at-risk adolescents. However, the computer-based assessment function has not been developed successfully: 'Because of substantive, technical and logistical issues that have delayed the development of an automated rating system, two trained researchers observed each adolescent's performance on the three VR vignette exercises' (Paschall et al., 2005, p. 64). The authors confirmed through email that this is still the case.

3.1.2 Assessment of similar constructs

Several methods appear similar to automatic social skill assessment, but they deliver other results (compare also Hasler, 2009, pp. 1-2):

- Often subsumed under the term 'social intelligence' are aspects of cognitive performance useful in social interactions, such as name memory or the decoding of nonverbal signs. The SJT has been validated as assessing these (e.g. Sternberg & Smith, 1985). These research advances can be clearly differentiated from the subject of this thesis, as they do not assess behaviour.
- Self-reporting social skills: since individuals typically see themselves differently from how others do (Greguras, Robie & Born, 2001), only low to moderate correlations can be found between self-ratings and ratings obtained from peers or supervisors (for a review, see Harris & Schaubroeck, 1988). Hence, self-rating in social skills should be perceived not as an assessment method for social skills but simply as what it is: a method of eliciting an opinion about a person's own skills.
- It is important to distinguish between social skills and personality traits, which are also manifested in social behaviour (Funder & Sneed, 1993). Social skills judgments have been found to correlate with personality, which has been interpreted in terms of the contribution of personality traits to an individual's ability to demonstrate socially competent behaviour (Ferris et al., 2001). They are different concepts, however. Klein et al. (2006) have pointed out that 'one can have acceptable social skills and still possess a deeply flawed personality'. The assessment re-

sults of an individual's social skills are likely to vary across measurement times because his/her competence may improve over time, with practice and feedback. On the other hand, personality tests are expected to provide the same results over multiple measurements of the same person, as personality traits are assumed to be stable over time.

- Social interaction analysis is often used in computer supported collaborative learning (e.g. see Resta & Laferrière, 2007, for an overview of many studies). However, this results in information about the social aspects of learning processes (to learn any subject), not about social skills.

Finally, many other personal characteristics that can be assessed automatically influence the performance of resolving social situations. In addition to intelligence, these include the following:

- Problem solving, planning, and organisational skills:
  Most non-multiple-choice assessment studies refer to computer-based business games and simulations (e.g. Drasgow & Olson-Buchanan, 1999; Drasgow, Olson & Keenan et al., 1993; Funke, 1998; Kleinmann & Strauss, 1998; Strauss & Kleinmann, 1995). Such simulations have been used in personnel selection and development since the 1970s. However, they are used for assessment of cognitive skills, such as complex problem solving or organisational planning skills. To date, these computer-based procedures have not yet been adapted for the assessment of social skills (Hertel, Konradt & Orlikowski, 2003).

3.1.3 Summary

As a summary of the above state of the art, no computer-based (or computerizable) tests have been successfully validated to assess social skills.

3.2 State of the art of social skill assessment relying on human judgment

3.2.1 Methods

Assessment centres and structured interviews are well-validated methods that are often used by companies to assess social skills (e.g. Kleinmann, 1997). These methods, however, require observation and scoring by human assessors thus making assessment expensive (see Section 2.2.1 for details). Social skills can also be assessed based on written free-text answers to questions about social situations, providing less but still significant validity (Funke & Schuler, 1998). These as well as methods utilizing everyday behavioural impressions in continuous work or educational relationships require no additional observation effort, but still need human judgment, thus causing expenses. Among the latter, validity is established for some well-designed 360-degree feedback instruments (which usually have social skills dimensions; e.g. Fletcher et al., 2002) and, in the case of school children, for rubrics for teacher assessment (e.g. Emerson et al., 1994) and for rubrics for assessment by parents (e.g. Merrell & Caldarella, 1999). Methods using everyday impressions, however, do not assess skill (performance at maximum motivation). They assess actual social behaviour as it is motivated (or not) by the circumstances.
In sum, while methods relying on human judgment can be highly valid (such as assessment centres), but the more valid they are, the more expenses they cause.

3.2.2 Issues of human assessment quality

Apart from practical concerns of expenses, there are also important issues of human assessment quality. While human assessors achieve (e.g. in assessment centres) satisfactory validity in determining the

overall skill of a person, this is only needed for the selection of job candidates (cf Section 2.2.3). For learning purposes, an overall judgment (e.g. 'has medium social skills') is obviously not sufficient. Educational assessment should deliver valid feedback regarding smaller entities of social skills. Unfortunately, even after decades of assessment research, the latter (called discriminant validity; Campbell & Fiske, 1959) is still an ongoing issue (e.g. Lance, 2008). Among others, the lack of discriminant validity is caused by typical human observation errors such as the halo effect, which is a cognitive bias where one trait of a subject influences the assessment of another trait. Computers, in contrast, are unlikely to exhibit such a bias.

A further problem related to the assessment of different aspects of behaviour is that the more aspects should be assessed, the more human assessors are needed. There is considerable evidence, that people are limited in their information-processing capacity (e.g. Baddeley & Weiskrantz, 1993). Consistent with current good practice (Ahmed, Payne & Whiddett, 1997; Gaugler & Thornton, 1989), the number of judgement dimensions assessed should be restricted to about four (Evers, 2005). Assessors' information-processing load is considerable, so that they must attend selectively to only a proportion of behaviours. Therefore, if they are overloaded with having to assess more dimensions, they may fail to perceive or misinterpret key behaviours, and may confuse the categorisation of behaviours by dimension (Fleenor, 1996; Gaugler & Thornton, 1989; Reilly, Henry & Smither, 1990; Thornton, 1992; Zedeck, 1986). It is self-evident that computers do not have these restrictions.

3.2.3 Summary

Assessment methods relying on human judgment perform well for selection of job candidates, albeit at considerable expenses. For educational purposes, there are basic issues of human information processing capability as well as biases, both of which computers do not have. Thus, there are gaps in assessment methodologies relying on human judgment, which call for an investigation if they can be filled by computers.

4 Design

This chapter presents the design of the computer-based assessment process. This design is described in the following sections:

- Section 4.2: The general, implementation-independent method for the assessment of social skills.
- Sections 4.3–4.7: The implementation of this method in a computer game, the PM Game (Project Management Game).
- Section 4.8: The concepts for creating stories and for steering stories and NPCs. Although these do not belong to the core assessment method, they are closely related as they influence the skills that can be assessed and the chances to extend the game.

In these sections, the motivations behind each design decision are provided. These motivations are not to be understood as scientific proofs of claims. Rather, they indicate that the design decisions have been made deliberately. Many of them are based on experiences with users testing the game. All of them were made in order to achieve concrete goals, namely, assessment validity and extensibility to further skills. A scientific investigation of the extent to which these goals were reached can be found in Chapter 5.

4.1 The PM Game

In order to facilitate an understanding of the abstract assessment method presented in the next section, the PM Game, in which the method is implemented, is first outlined. The terms used for the basic components of the game are defined. The intention behind this is to enable non-game researchers to understand the rest of this thesis.

For a visual introduction to the game, see Figure 46 to Figure 50 in Appendix 7.6. These figures show a sequence of screenshots taken while playing the game. For screenshots of the authoring interface, see Figure 42 to Figure 45.

The PM Game is a single player game that permits the user to communicate with simulated people (‘non-player characters’ or NPCs). One playable unit of the game is called a ‘story’ (see Appendix 7.1 for the currently available stories). There is no personalisation or persistence in the game. That is, each story always behaves identically, independent of the players and how often they play it. The game is driven by only mouse clicks (that is, it is a ‘point & click’ game). In order to communicate with an NPC, the user creates sentences using a menu. Typically, 3 to 5 clicks are required to create one user communication. NPCs answer automatically in the form of a speech bubble. In contrast to most games, there is no concept of space or movement, making this purely a communication game. Although one part of the interface resembles a map and there are locations on it, no players, NPCs, or objects can be located in these places. The user and the NPCs only talk about the locations.

Many additional components and concepts of the PM Game are introduced in Sections 4.3 to 4.8. Detailed information on the software architecture, as well as on Web 2.0 and the educational aspects of the PM Game, can be found in the thesis by Ito (2009).

4.2 The assessment method

The assessment method proposed in this thesis is a procedure consisting of the following parts:

Part 1. Present to the user at the same time:
   a. a social situation (consisting of elements such as communication partners, third persons, objects, locations, abstract concepts relevant to the situation, attributes of all the aforementioned points, actions the user can take, and past communications)
   b. a fixed menu for creating verbal communications. This menu is independent of the social situation (it contains sentence stubs such as 'What do you think about … ?')

Part 2. Make the user generate communications intended for simulated persons (NPCs) by letting the user combine elements of the menu with elements of the social situation to form complete sentences (for example 'What do you think about Thomas?'). The menu elements and elements of the social situation can be freely combined as far as grammatical rules permit.

Part 3. Scan the resulting interactions between the user and the game for information that is indicative of a specific social skill (for example, providing summaries is an indicator of good meeting moderation abilities).

Part 4. Compute skill scores from the observations. (Here various algorithms can be used, from simple addition to more complex algorithms relating the produced behaviours to opportunities and weighing the impact that an observed behaviour may have on performance, etc.).

The above description is general in order to emphasize the idea of the method; and it is example-based to promote a better understanding. This general method can be used in different arrangements and with different implementations of the parts.

In this thesis, the following arrangement is investigated: Parts 1 and 2 constitute the playable game. The actions taken by the user (Part 2) result in changes in the social situation (Part 1a). Parts 3 and 4 constitute the scoring mechanism, which is initiated after playing a story. The result is a score for the skill of mastering a particular task in the story.

Another arrangement may be to present a social situation without updating it. That is, there is no game as such, but a single situation is presented to the user. In this case, the test would consist of producing one single reaction. This test arrangement could be of interest as it has been used successfully by Funke & Schuler (1998). They used this arrangement with textual situational descriptions, asking the test takers to produce a written free text reaction to each situation. Social skills were assessed using trained human judgment.

The concrete implementation of each part of the method used in the PM Game will be described in Sections 4.3 to 4.7.

Summary and motivations behind design decisions

As a summary, the core idea is to separate the presentation of the current situation (Part 1a) from the user interface that provides the menu-based communication via a fixed menu (Part 1b). By combining these two, ready communications are generated that refer to the concrete social situation (Part 2). This solution was chosen in order to simultaneously achieve the following:

1. Limit what the user can talk about to communications that can be judged semantically by the computer. Assessment of communicative skills requires a semantic judgment of the communications. This is realized by the above design, as the elements of the menu are fixed and situation independent, and the elements of the social situation that the user can talk about are limited to what the game presents to the user. Hence, the pool of possible communications that the user can generate is large but can be checked in advance.
   Remark: It is nonetheless possible that unexpected semantics or unexpected situations could occur as a result of specific sequences of communications by the user. These must be handled using several story design measures (see Section 4.8).
   Essentially, this design prevents out-of-context, nonsense, or provocative communications, which would be possible with natural language. Such communications would not contribute to solving the presented social situations. They are difficult to handle properly without human intelligence or sophisticated artificial intelligence. Experience with the PM Game and other communication games such as Façade shows that users immediately start looking for opportunities to produce such communications.
   The decision to limit what the user can talk about does not come without a cost. For example, there are limitations on modelling small talk. By nature, small talk, or 'phatic conversation' (Malinkowski, 1923), is not usually limited to the concrete situation that will be discussed afterwards. However, the ability to make small talk is a useful skill, e.g. to deduce the social position of the other person (Laver, 1975). Thus, it is a loss that this assessment method is not able to model it properly.

2. Despite point (1), the users should not be limited excessively in expressing what they want to do in the given situation. If the users cannot say what they want, it will be impossible to observe their real social behaviour. This requirement is fulfilled by the above design because the menu and situation elements can be freely combined in any manner that grammatical rules permit. This can add up to several hundred possible communications in a complex game situation.
   Remark: However, it is possible, and in fact common, that during the testing of stories, users want to say something important but cannot. Once again, this needs to be handled by the story design measures (see Section 4.8). Experience shows that, with some limitations on the possible story content, all important user communication intentions can be expressed.

3. The possibilities for what sentences can be formed should not change with the situation because users would spend a substantial amount of time browsing through options to explore the possible sentence for a given concrete situation. This has been the actual experience with prototypes of the PM Game, implying that the thinking processes that are induced are very different from those in real social situations, where the means of linguistic expression remain the same and people do not browse through predefined alternatives of action but have to develop action alternatives themselves.
   Remark: Even with the proposed method, at least in the concrete implementation of the PM Game, users can be observed to browse through the communication menu. However, they eventually realize that the same options are always offered, and they can rely on their constant availability

It should again be pointed out that the above arguments are not to be understood as proof of any scientific claim, but are merely an explanation of the choices made in the design of the method as a means of solving problems (remark: further motivations behind the selection of individual parts of the meth-

od can be found in the next few sections). The scientific argument explaining why the separation of the situation and the situation independent communication possibilities is a central design element of the method can be found in Section 6.4. This was essentially done to avoid presenting social judgment tests to the user (multiple-choice tests that present ready action alternatives to users, including both the situation elements and the sentence). Based on meta-analytic evidence, these are not capable of assessing social skills. A comparison of the mental processes involved in real social situations when solving social judgment tests versus playing the PM Game would be an interesting topic for further research (see Section 6.3).

4.3 Part 1a: Presentation of social situations

In order to assess social skills, the PM Game presents social situations to the user. The presentation is simple: there is no real-time action, no movement, and only 2D graphics are used. The social situations evolve turn-wise in response to the actions of the user. In a story, the presentation of the situation can currently have the following parts. Each part consists of presenting simple but suggestive icons and/or short texts to the user.

Introductory text

At the beginning of a story, a piece of text is displayed in the left-side bottom corner. This provides the user with information about his or her role and the mission to be completed (Figure 3). In parallel, the user can observe elements of the situation referred to in the text.

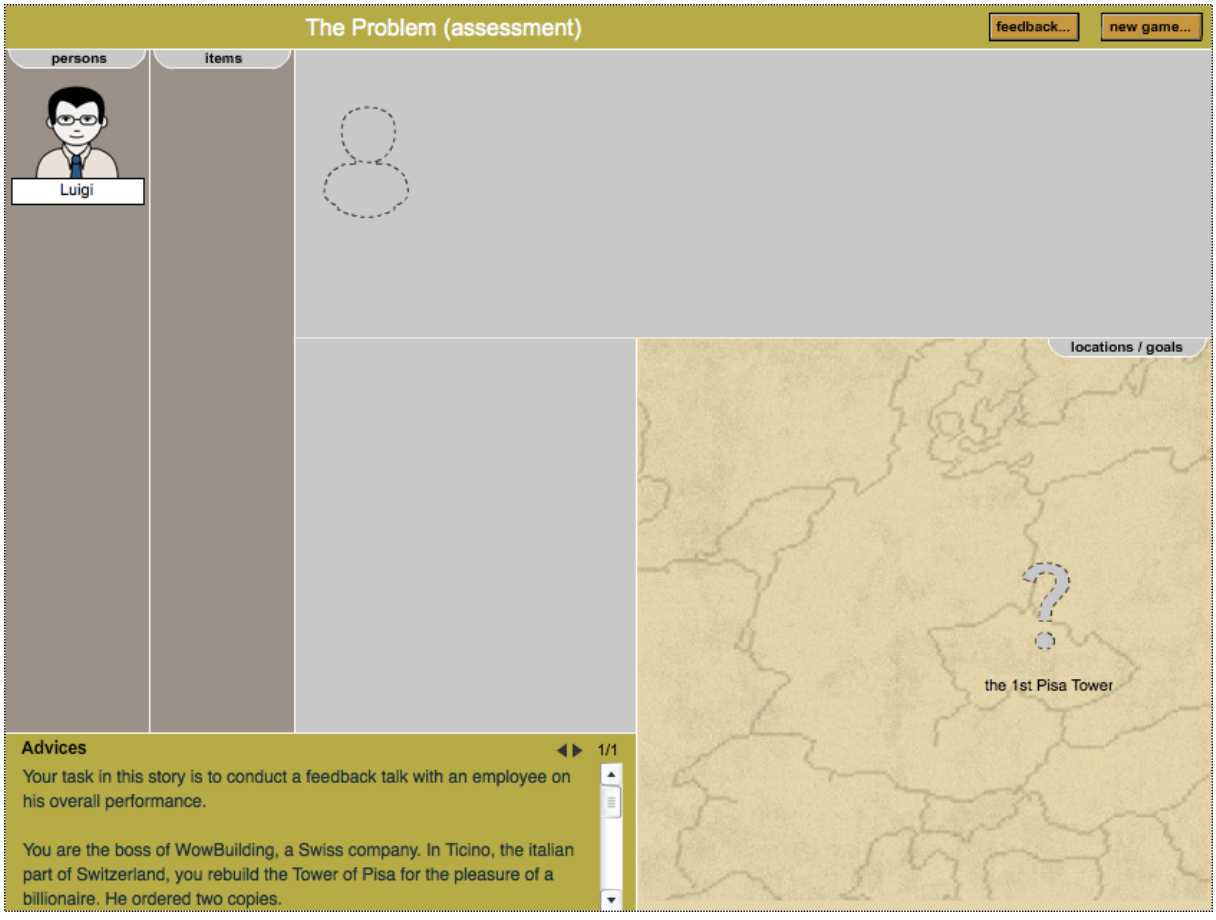


Figure 3: Screen at the beginning of the story ‘The Problem’

Elements of the social situation

The elements of the current situation are always presented to the user in the form of a collection of icons (small pictures with captions) arranged on the game interface. Technically, there are three different kinds of icons: persons (NPCs one can talk with), locations (objects presented on a map, e.g. a house), and items (objects presented in a list). These can be used in a variety of creative ways: abstract concepts such as ‘safety’ or ‘holidays’ are displayed as items (Figure 4). In the story ‘Mary and Garry’, an event is depicted by an item entitled ‘the wedding’. Further, an item is used in the story ‘The Problem’ to display a third person (Figure 4). Technically, this is an item, and not an NPC that the user could talk with. However, it looks like a person. Thus, it functions as a ‘third person’ about whom the user can talk.

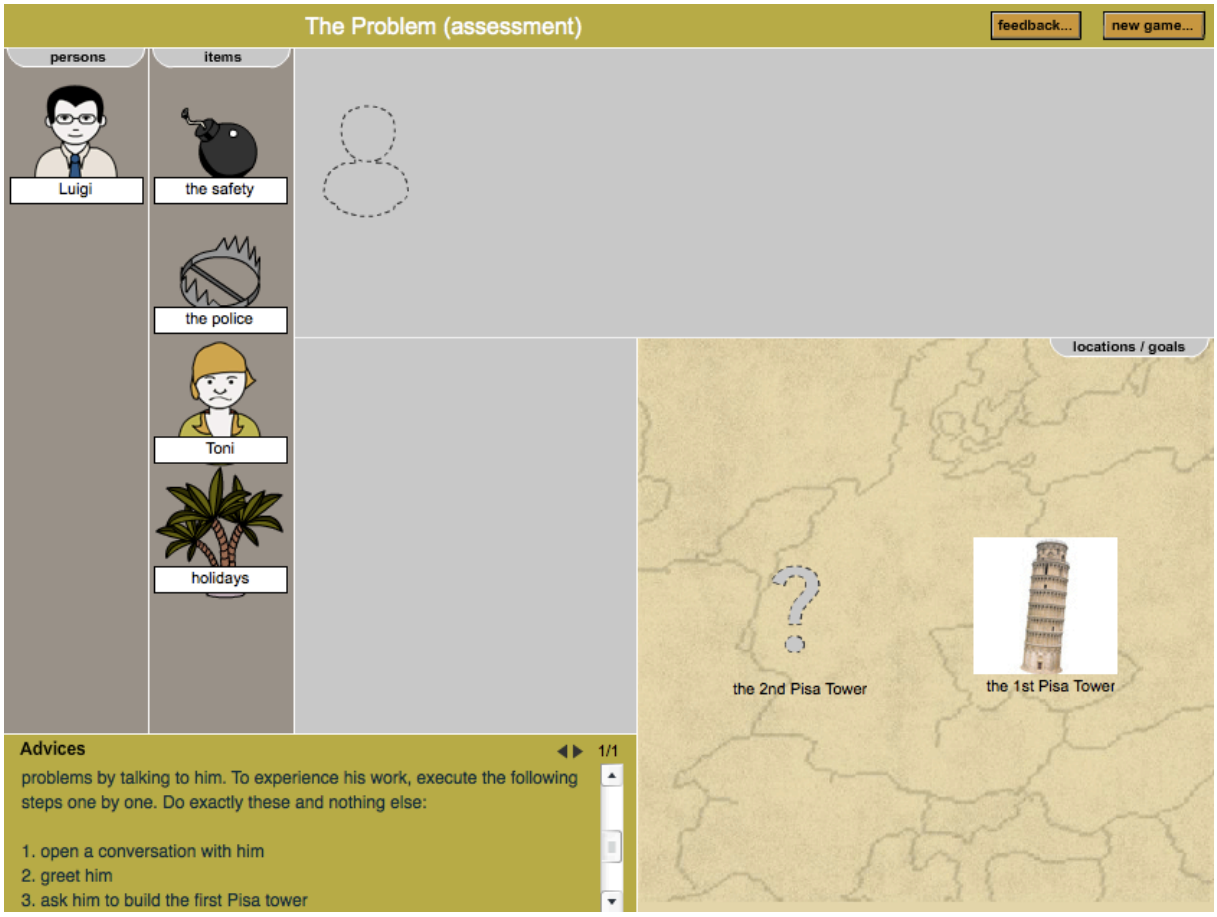


Figure 4: Icons representing parts of the situation

Details pertaining to situation elements

When the mouse is placed over an icon, details become visible (Figure 5). These include actions that the user can perform related to the object, attributes the user can talk about, and some descriptive text.

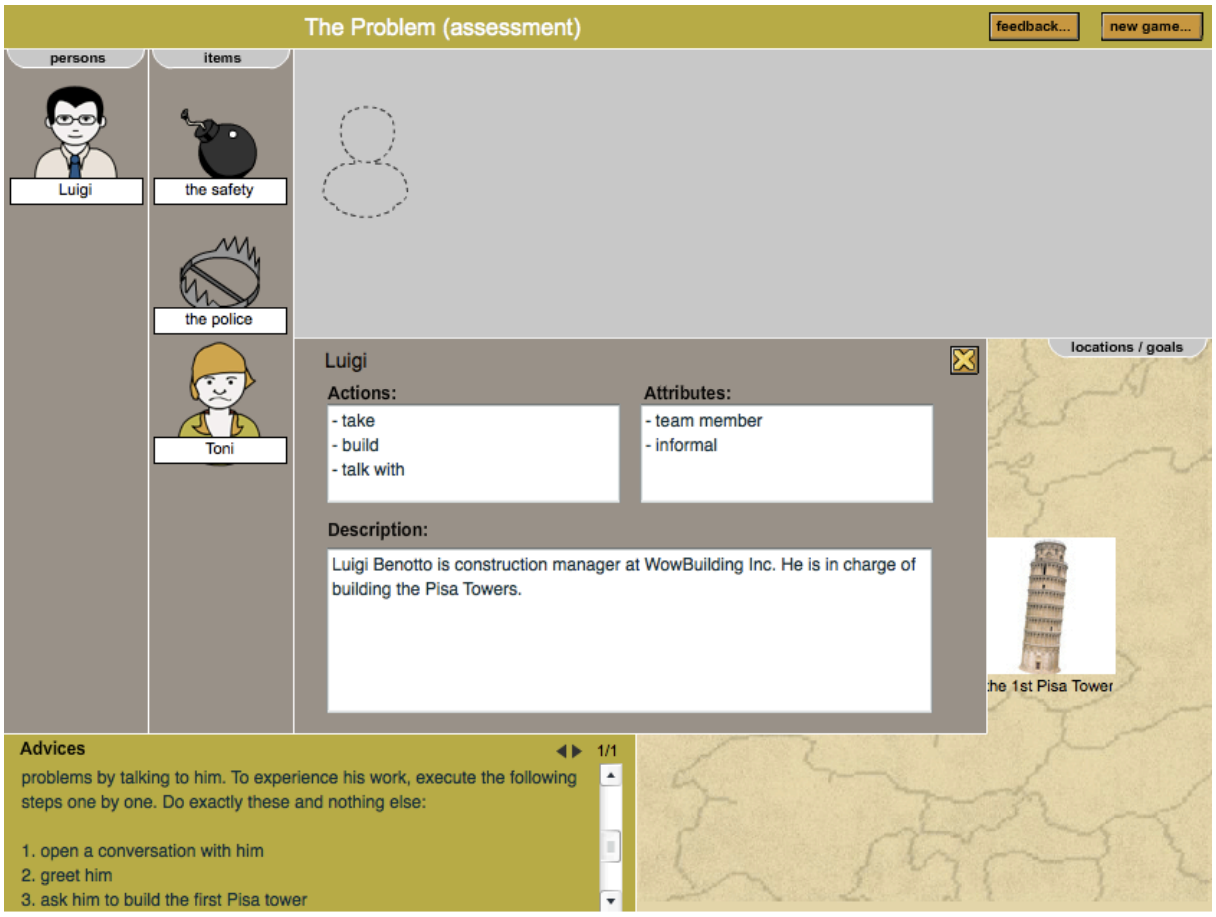


Figure 5: View showing details of the NPC Luigi in the story ‘The Problem’

Current dialogue situation

The icon of the conversation partner (NPC) and the dashed-line representation of the user, along with the speech bubbles between them, show the current moment of the conversation. The faces of the NPCs are capable of the following simple emotional expressions: very happy, happy, neutral, sad, and very sad (compare Figure 6, in which Luigi smiles, with Figure 9, in which Luigi is neutral).

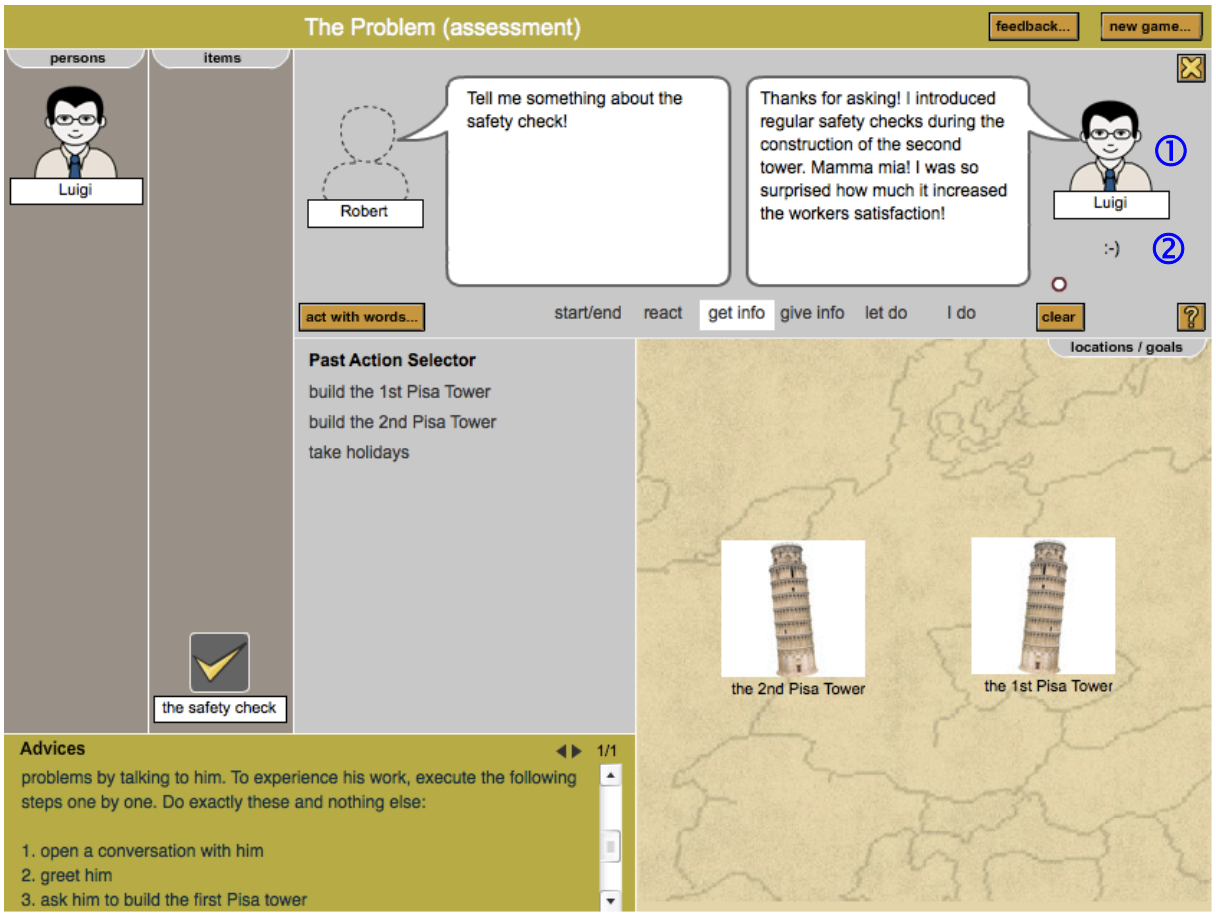


Figure 6: Luigi smiling ((1) and (2))

Experiencing the background story leading up to the situation

By combining the above elements to display social situations, it is possible to let the user participate in the events leading up to the situation. The purpose of this is to provide information to the user in a phased manner, in order to avoid overwhelming the user by explaining all the details at once. First, when the user starts the game story, only the role of the user and the environment in which the user acts are explained briefly (Figure 3). Then, the user is instructed to say a specific thing to a specific conversation partner, for example to delegate a task to an employee (Figure 7).

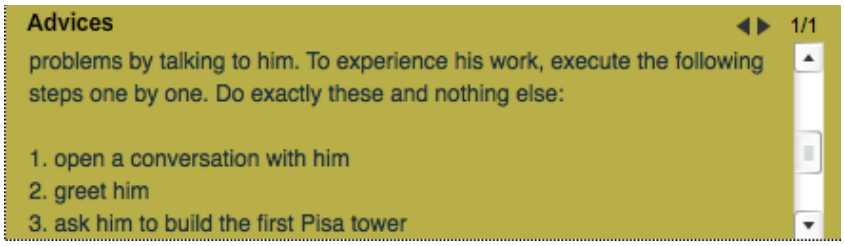


Figure 7: Instructions to the user for receiving the next piece of information

The events caused by the user's communication result in a changed situation, which is the real starting point of the story. For example, something went wrong when an employee carried out a task. The goal of the story is now explained; for example, the goal might be for the user to give feedback to the employee to prevent a reoccurrence of the problem (Figure 8).

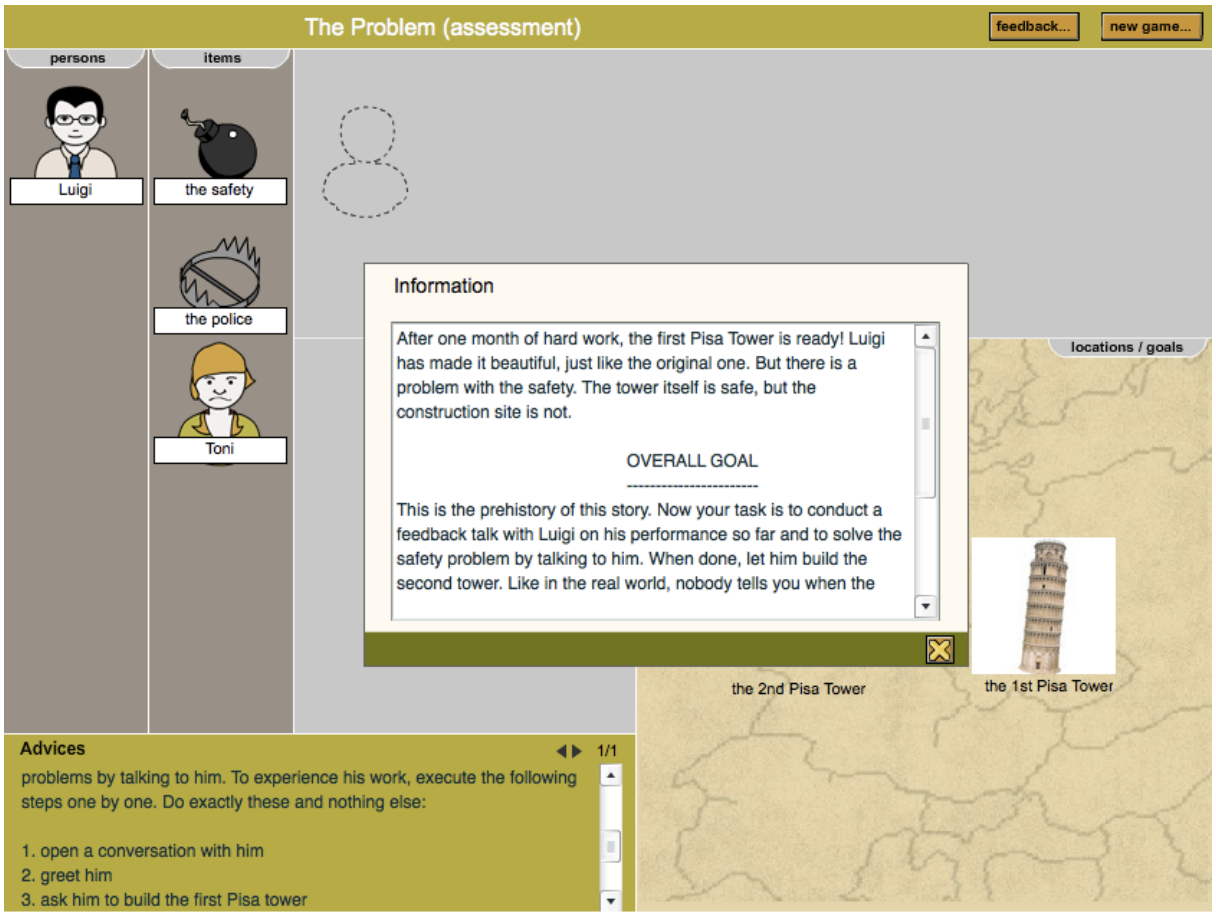


Figure 8: Real starting point of the story ‘The Problem’

The past

There is a box on the game screen called ‘past actions’ that lists the actions that have already occurred in the current story (Figure 9). As more actions are performed, the list is lengthened (Figure 6). At the current stage in the development of the PM Game, this is the only aspect of the past of the game story that is represented on the screen.

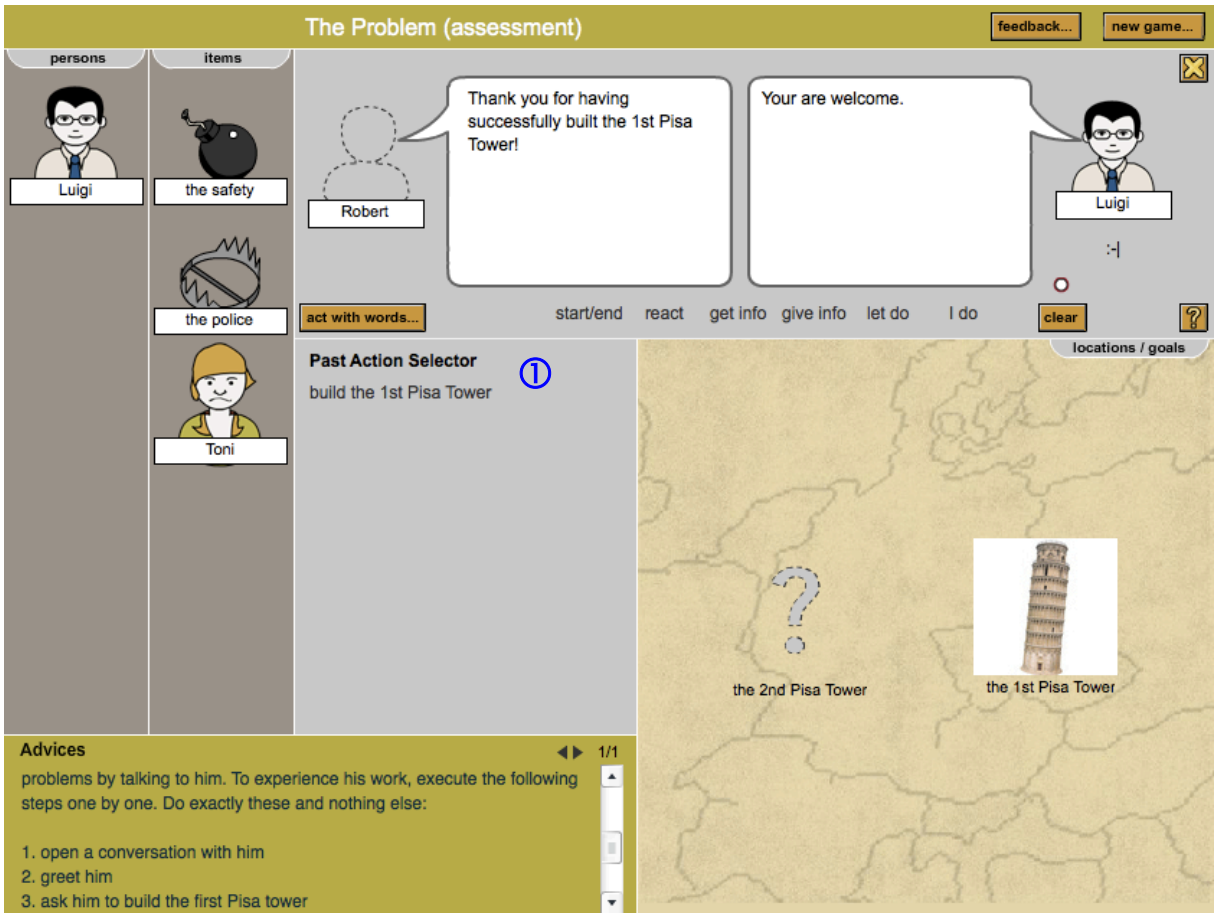


Figure 9: Display of past actions (1) as a part of the social situation

Summary and motivations behind design decisions

In sum, the design of the presentation of the social situation by the PM Game adheres to the following principles:

- Avoid anything that could distract the user from the communicative purpose of the game (such as bodily movements or 3D graphics).
- Avoid overwhelming the user with too much information at once. Information is presented in small chunks on the screen, as well as bit-by-bit as a social situation evolves.

The motivations for these design principles are to optimise for assessment validity and extensibility of the PM Game:

- The cognitive load of the users should be minimized. Exposing users to a cognitive load may result in an assessment of the cognitive performance instead of social skills (see Section 5.3.7 for a definition of cognitive load and the social skill assessment problems associated with it).
- It has been shown, both empirically and theoretically, that the response format is more relevant than the stimulus format for assessment purposes (Karkoschka, 1997, p. 203ff; Funke & Schuler, 1998). In other words, the stimulus format (that is, the presentation of the social situations) can be simple. This insight from assessment psychology can readily be used to design an extensible IT system because it represents a justification for avoiding all aspects of state-of-the-art representations of social situations such as 3D graphics, sound and movement. These would typically increase the effort needed to extend the system with new stories because even with the use of efficient tools and libraries to design a 3D world with as little effort as possi-

ble, there is always a need to specify where to position NPCs and other objects in space and how they should move as a part of the social situation.

- Observing users of the PM Game has shown that distraction is a major pitfall. For example, students attempted to make the customer in the story 'Mary and Garry' drunk by repeatedly offering a drink. This indicates that, even the limited possibilities in the game for playful behaviour, irrelevant to the assessment, are actively exploited, justifying the general reduction of such possibilities in the game.

In view of the simplicity principle regarding the appearance of the game, it can be questioned if the icons could be avoided in favour of a purely textual description. However, regarding extensibility there would be no substantial gain. The game contains a library of pictures. Experience shows that with this it is usually possible to cover the needs imposed by new stories. This is best demonstrated by the fact that almost all pictures used by the three stories were originally prepared for other stories. Regarding possible distraction of the users by the pictures no effects have been found in observations of individual users, however this question has not been systematically investigated.

4.4 Part 1b: Situation-independent communication menu

As introduced in Section 4.2, the main vehicle of user communication in the game is the communication menu. This menu contains sentence stubs. These are sentences with gaps, where each gap has a specification of the kind of game object that can be used to fill it.

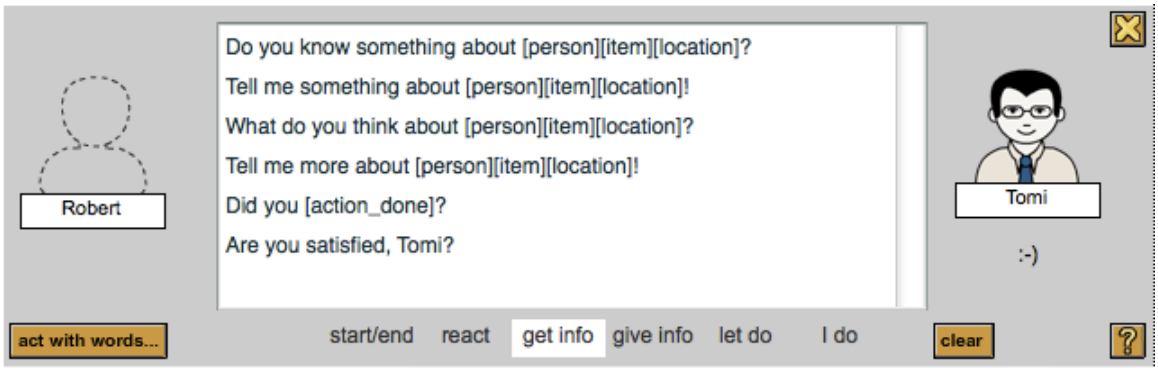


Figure 10: Sentence stubs under the menu entry 'get info'

As an example, the sentence stub 'What do you think about [person] [item] [location]?' can be used to produce the sentence 'What do you think about Michael?' (see Section 4.5 for the mechanism used to speak). The sentence stubs are grouped into the following categories of the main menu:

start/end react get info give info let do I do

As an example, Figure 10 shows the sentence stubs available when clicking on the menu entry 'get info'. Both the groups and the sentence stubs are the same in every story of the PM Game. For a complete list of all the sentence stubs in all the menu entries, see Appendix 7.5.

There is a distinction regarding the level of formality of the conversation partner: each NPC has an attribute is_formal, which can take the values 'formal' and 'informal'. Depending on this attribute, a more or less formal version of the same sentence stub can be displayed. As an example, Tomi in Figure 10 above had the is_formal attribute set to 'informal'. Figure 11 displays the same sentence stub

list in the case where is_formal is set to 'formal'. An example of the difference can be seen in the sentence 'Were you able to [action_done]?' (e.g. 'Were you able to pay the bill?').

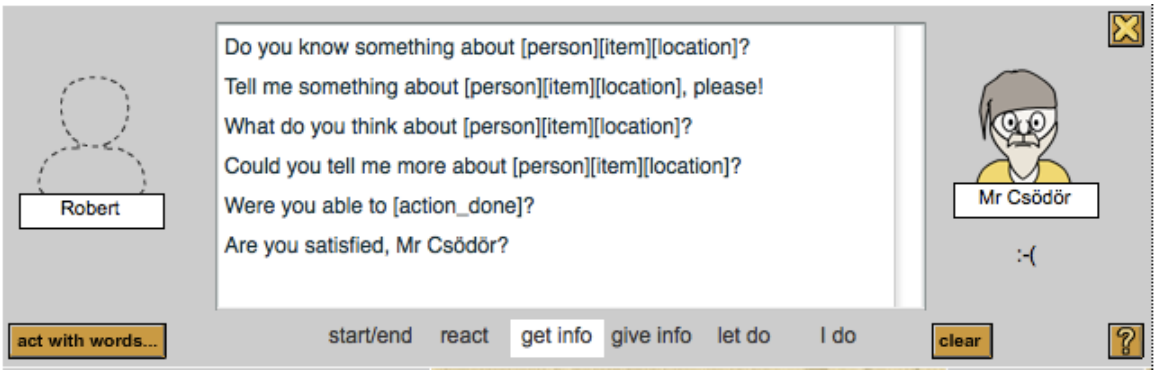


Figure 11: Formal sentence stubs under the menu entry 'get info'

Motivations behind design decisions

The above design was chosen to be as simple as possible while satisfying the needs of the assessment method proposed in Section 4.2: 'Part 1b: a fixed menu to create verbal communications'. This menu is independent of the social situation. However, there might be issues regarding the extensibility: although a limited set of sentence stubs may be sufficient to cover selected situations, issues might arise when the game needs to assess additional skills, which requires the addition of more and more situations. This threat to extensibility will be discussed in detail in Section 5.6.

4.5 Part 2: Combine menu and situation elements to produce sentences

The game has two basic states: in a conversation or outside a conversation. Outside a conversation, the user can look at the elements of the social situation (details become visible on mouse-over) or drag one or several NPCs into the communication interface to initiate a conversation. When the user is in a conversation, the main menu described in the previous section becomes visible. To create a sentence intended for an NPC, the user selects one of the following in any order:

- A category (start/end, react, etc.) from the main menu (Figure 12)
- Any element (person, item, location, or past action) of the social situation visible on the screen (Figure 13).

Usually, there are several sentences matching a category from the main menu and an element of the social situation. Thus, a further click is required to select a sentence (Figure 14). Finally, the user 'says' the resulting complete sentence by pressing the checkmark button (Figure 15). All this happens without time pressure: the PM Game is turn-based and waits until the user has pressed the checkmark button. The user only sees the results of his action after pressing this button (Figure 16).

There are special cases: some sentence stubs have no gaps. These can be found only by clicking on a category in the main menu. Other sentence stubs have gaps for attributes, for example 'I am *feeling*.' Here, the '*' sign indicates that this is not an ordinary gap to be filled with an element of a social situation. When clicking on this sentence, the possibilities that can be used to fill this special gap become visible, in this case 'happy', 'sorry', 'sad', etc.

Summary and motivations behind design decisions

As a summary of the above, the user communication in the PM Game:

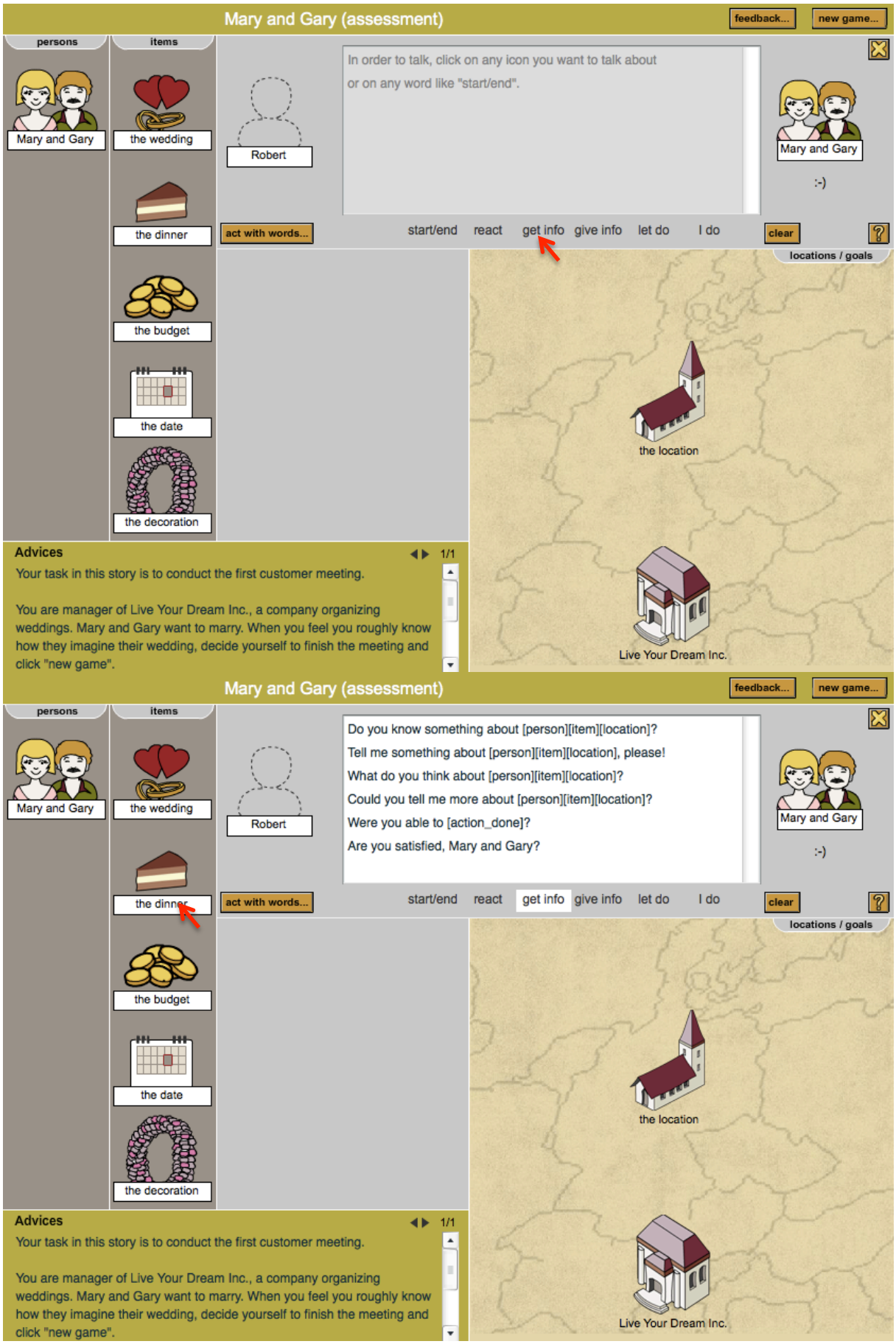
- Occurs in a turn-based manner, without time pressure,
- Follows a simple design principle: the user can click on anything visible to talk about it,
- Confronts the user with tags such as [item] or *feeling*.

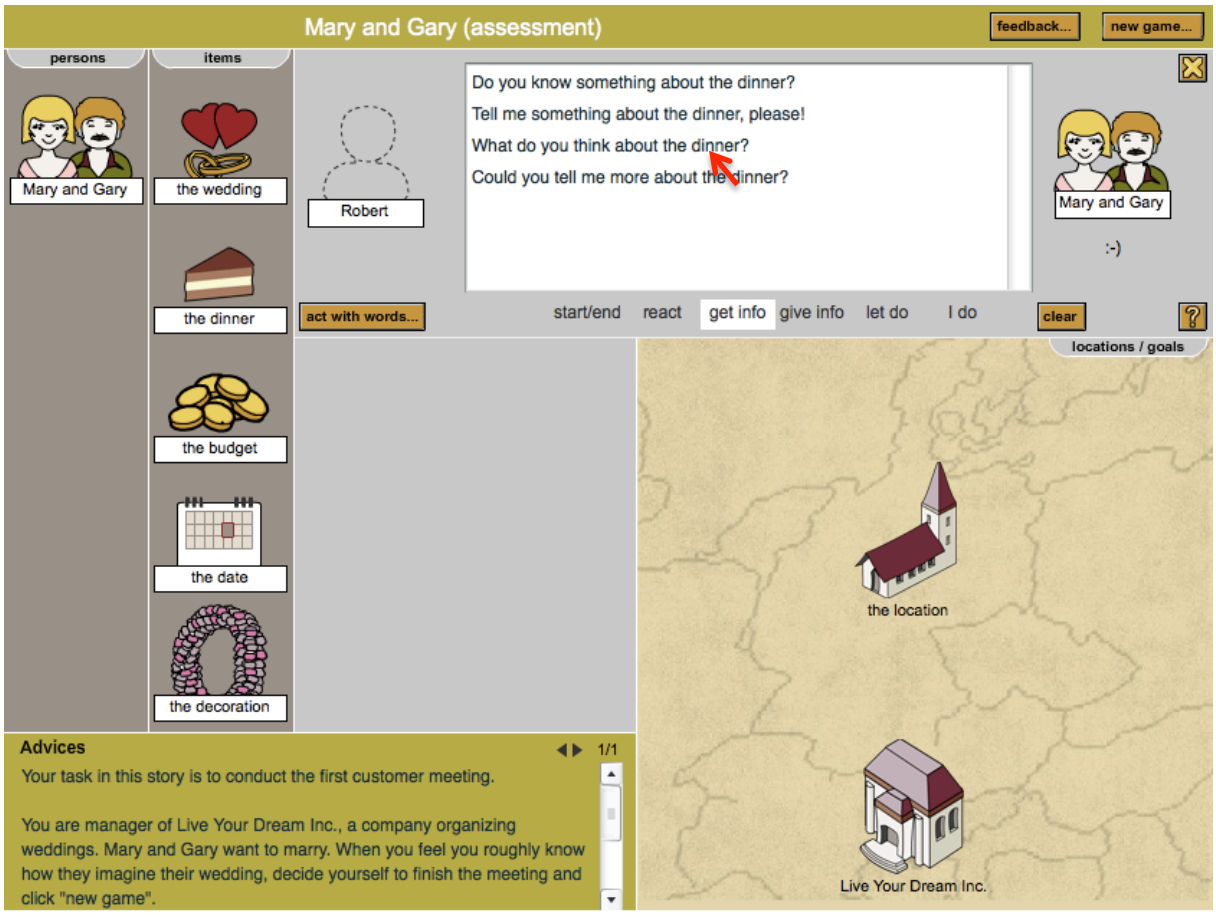
Motivations:

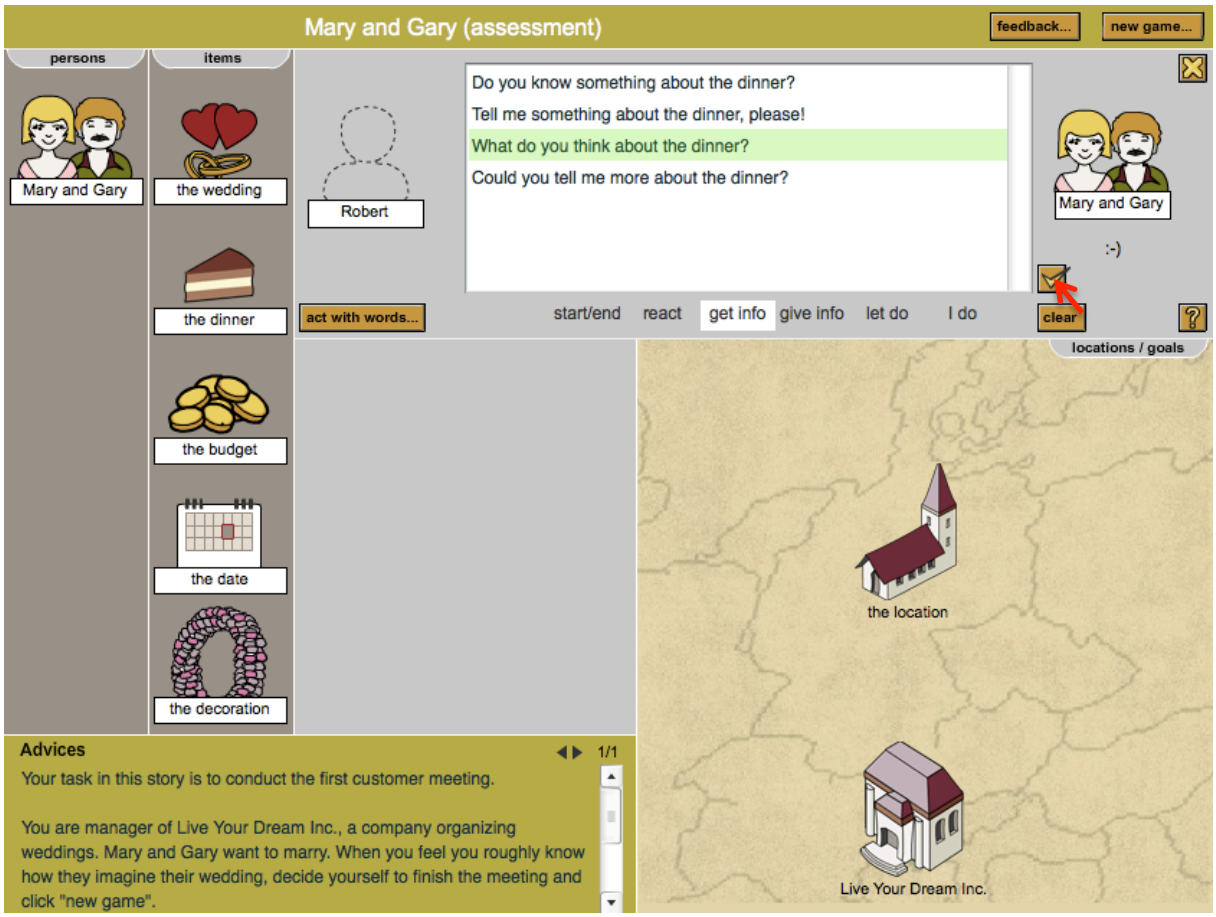
- No time pressure: pretests with users have shown that people need greatly varying amounts of time to familiarize themselves with the PM Game. Thus, it has been concluded that the interface represents differing levels of challenge to users. Hence, it has been estimated that time pressure would be likely to put users under pressure regarding their use of the interface in the assessment stories. Thus, there would be a risk of assessing the skill needed to quickly cope with this unnatural interface instead of assessing social skills.
- Click-on-anything design: observations of users indicate that this is appreciated.
- Tags: although users have confirmed that the tags do not disturb their understanding of the PM Game, it is currently unknown whether their meaning is understood and whether they represent an advantage or a disadvantage for the majority of users.

Screenshots

For better visibility, a red arrow is used to indicate the mouse position.


Mary and Gary (assessment)
feedback...
new game...
persons
items
Mary and Gary
the wedding
Robert
In order to talk, click on any icon you want to talk about
or on any word like "start/end".
Mary and Gary
:-)
the dinner
act with words...
start/end
react
get info
give info
let do
I do
clear
locations / goals
the budget
the date
the location
the decoration
Advices
1/1
Your task in this story is to conduct the first customer meeting.
You are manager of Live Your Dream Inc., a company organizing weddings. Mary and Gary want to marry. When you feel you roughly know how they imagine their wedding, decide yourself to finish the meeting and click "new game".
Live Your Dream Inc.
Mary and Gary (assessment)
feedback...
new game...
persons
items
Mary and Gary
the wedding
Robert
Do you know something about [person][item][location]?
Tell me something about [person][item][location], please!
What do you think about [person][item][location]?
Could you tell me more about [person][item][location]?
Were you able to [action_done]?
Are you satisfied, Mary and Gary?
Mary and Gary
:-)
the dinner
act with words...
start/end
react
get info
give info
let do
I do
clear
locations / goals
the budget
the date
the location
the decoration
Advices
1/1
Your task in this story is to conduct the first customer meeting.
You are manager of Live Your Dream Inc., a company organizing weddings. Mary and Gary want to marry. When you feel you roughly know how they imagine their wedding, decide yourself to finish the meeting and click "new game".
Live Your Dream Inc.


Mary and Gary (assessment)
feedback...
new game...
persons
items
Mary and Gary
the wedding
the dinner
the budget
the date
the decoration
Robert
Do you know something about the dinner?
Tell me something about the dinner, please!
What do you think about the dinner?
Could you tell me more about the dinner?
Mary and Gary
:-)
act with words...
start/end
react
get info
give info
let do
I do
clear
locations / goals
the location
Live Your Dream Inc.
Advices
1/1
Your task in this story is to conduct the first customer meeting.
You are manager of Live Your Dream Inc., a company organizing weddings. Mary and Gary want to marry. When you feel you roughly know how they imagine their wedding, decide yourself to finish the meeting and click "new game".


Mary and Gary (assessment)
feedback...
new game...
persons
items
Mary and Gary
the wedding
the dinner
the budget
the date
the decoration
Robert
Do you know something about the dinner?
Tell me something about the dinner, please!
What do you think about the dinner?
Could you tell me more about the dinner?
Mary and Gary
:-)
act with words...
start/end
react
get info
give info
let do
I do
clear
locations / goals
the location
Live Your Dream Inc.
Advices
1/1
Your task in this story is to conduct the first customer meeting.
You are manager of Live Your Dream Inc., a company organizing weddings. Mary and Gary want to marry. When you feel you roughly know how they imagine their wedding, decide yourself to finish the meeting and click "new game".

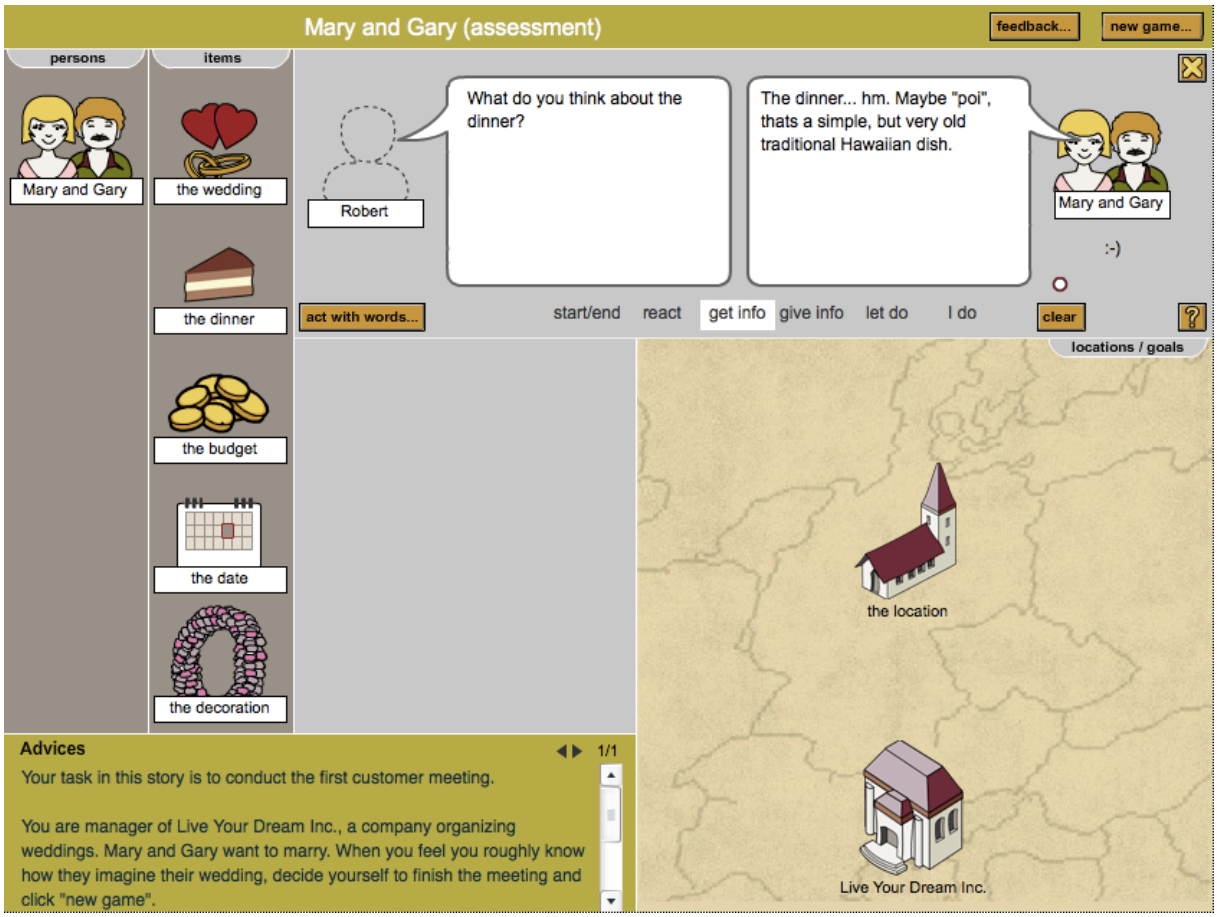


Figure 12: Click ‘get info’: Click 1

Figure 13: Click ‘the dinner’: Click 2

Figure 14: Click on a sentence: Click 3

Figure 15: Click on the checkmark button: Click 4

Figure 16: Resulting user sentence and answer from the NPC

4.6 Part 3: Behavioural markers

In order to score skills, the PM Game observes behaviour. It identifies the presence or absence of behavioural markers. In this section the scientific concept of behavioural markers will be introduced; it will be needed in Section 4.7 to describe the scoring algorithm.

Behavioural markers are defined as ‘observable, non-technical behaviours that contribute to superior or substandard performance within a work environment’ (Klampfer et al. 2001, p. 10). A good behavioural marker is characterized as follows:

- It describes a specific, observable behaviour, not an attitude or personality trait, with a clear definition (the enactment of skills or knowledge is shown in a behaviour).
- It has demonstrated a causal relationship to the performance outcome.
    - It does not have to be present in all situations.
    - Its appropriateness depends on the context.
- It uses domain specific language that reflects the operational environment.

- It employs simple phraseology.
- It describes a clear concept.

For the PM Game, a particularly important part of this is that the appropriateness of a behavioural marker depends on the context. For example, providing a summary is a typical indicator of professional meeting moderation. Yet, it is only meaningful if there is something to summarize, that is, if several pieces of information have been received. For this reason, the behavioural markers in the PM Game are story specific. They must be crafted for the specific story, along with the story content. In this regard, there is no difference compared to the role-plays used in traditional assessment centres for judging social skills: both the scripts for the role-plays and the scoring criteria used by the human assessors are created in advance.

A further important part of the characteristics of ‘good behavioural markers’ cited above is that they ‘do not have to be present in all situations’. For example, providing a summary in a customer meeting is usually a professional behaviour that helps the task of understanding the requirements of the customer (allowing the customer to identify a deficiency or approve the information stated). However, it is not mandatory to produce a summary when a customer expresses requirements. Moreover, there are other means to resolve the same need. For example, after a customer makes a statement, it can be repeated or rephrased to allow the customer to immediately correct the speaker’s understanding, if necessary.

Behavioural markers in the PM Game are designed to adhere to the above definition and quality criteria. However, in the following, it will be described that there are differences between the usage of behavioural markers by professional human assessors and the usage in the PM Game. These differences affect the typical 'size' of the behaviours and the rating methods used.

Behavioural markers for human assessors tend to be larger, more complex behaviours such as ‘Discusses case with surgeons or colleagues’ (Fletcher et al., 2004), which is a complex series of interactions. The behavioural markers of the PM Game tend to be smaller, such as ‘provide a summary’. Thus, a typical behavioural marker for usage by human assessors may correspond to a multitude of different behavioural markers used by the PM Game (a summary can be a small part of a discussion).

Human observers usually rate behavioural markers on a Likert scale with several scale values such as 'good', 'acceptable', 'marginal', 'poor', and 'not observable' (e.g. Fletcher et al., 2004). That is, the judgment involves simultaneously a statement on presence of a specified behaviour as well as on its quality. Compared to this, the PM Game rates the presence (or absence) of behaviours relative to how many times the behaviour would have been useful. For this, the following two numbers are computed for each behavioural marker:

f	Number of occurrences of the behaviour (f like 'fulfilments')
p	Number how many times the behaviour can be expected to have occurred in the dialogue for a good performance (p like 'oPportunities')

Table 4: The PM Game 'rates' behavioural markers by counting fulfilments and opportunities

This is a simplification as even a 'small' behavioural marker such as a summary may be performed in different qualities (accuracy, completeness, etc.); this is however ignored in the PM Game. In fact, even an incomplete summary counts as a summary.

The present section serves to introduce the concept of behavioural markers. How the above numbers will be computed in different cases will be specified as part of the scoring algorithm in Section 4.7. An example of a user-NPC dialogue with all the behavioural markers identified can be found in Appendix 7.4.3. The complete list of behavioural markers used in the PM Game can be found on the accompanying CD[1].

Summary and motivations behind design decisions

Behavioural markers are an established concept for judging human behaviour. The PM Game adopts behavioural markers to judge behaviour, however changes their typical 'size' and rating method. Essentially, the idea is to shrink the complexity of the behaviours so they can be considered as either being 'present' or 'not present' in a binary fashion. The goal behind this design decision is to avoid the need for human common sense and situation experience to decide what a ‘poor’ or ‘acceptable’ performance of a behavioural marker is.

Remark: the simplicity of 'present' or 'not present' does not render the automatic judgment by the PM Game context-insensitive, as will be shown by several examples in the next section. The solution is to code the context into the functions to compute the number of fulfilments and the number of opportunities for a behavioural marker. That is, these functions may contain conditions referring to the context of the behaviour. Under these conditions a fulfilment or an opportunity is counted.

4.7 Part 4: Scoring algorithm

Each score a user receives for a particular skill is computed using their performance in a single story. While the user plays the story, the PM Game writes a log file containing the dialogue between the user and the NPC. In the following, the algorithm is described to extract skill scores from the log file.

Algorithm structure

The algorithm consists of the following parts:

- For each behavioural marker i belonging to a skill k, there is a specific algorithm to extract the number of fulfilments and the number of opportunities from the log file (see the end of Section 4.6 for a definition of fulfilments and opportunities)
- Using the above data, a score aggregation algorithm computes the score for the skill k

Score aggregation algorithm

After the user has played a story, the following algorithm computes the score for a skill k.

```
float function score (int k) {
      float x = 0;
      for (int i = 0, i++, i < nk) if (pki != 0) then x += fki / pki;
      int y = 0;
      for (i = 0, i++, i < nk) if (pki != 0) then y++;
      if (y != 0) return x/y else return -1;
            // -1 for 'undefined'
}
```

[1] File names ’M&G - assessment by PM Game rules.xls’ and ‘TheProblem - assessment by PM Game rules.xls’

The above algorithm makes reference to several variables that are to be defined and filled with the appropriate data:

- int n_k : number of behavioural markers used for assessing the skill k
- int f_{ki} : number of times the user had fulfilled behavioural marker i in skill k
- int p_{ki} : number of opportunities the user received to fulfill behavioural marker i in skill k

Remark: For n_k, f_{ki} and p_{ki}, the more compact index notation was chosen instead of the array-notation (`n[k]`, `f[k][i], p[k][i]`) in order to improve readability of the following explanations.

In sum, the scoring algorithm calculates the non-weighted average of the fulfilment versus opportunity ratios of the individual behavioural markers belonging to the skill that is to be assessed. In addition, the algorithm takes into account that, depending on the specific course of action while playing a story, it might happen that a user received no opportunities to produce a behavioural marker.

Calculations specific for each behavioural marker

While n_k is a predefined constant value, f_{ki} and p_{ki} are computed using functions operating on the log file in which the PM Game has protocolled the dialogue between the user and the NPC. These functions are specific for the behavioural marker i of the skill k; they contain the behavioural knowledge how user behaviour in the game is to be scored. As the number of these functions is substantial, they are provided in the form of Excel sheets on the accompanying CD[2]. In the following, the functions for computing f_{ki} and p_{ki} will be described on a conceptual level. To facilitate understanding, behavioural markers will be divided into for groups A, B, C and D based on how f_{ki} and p_{ki} are calculated. In groups A and C, p_{ki} is simply a constant value for a given k and i. In groups B and D, p_{ki} is dependent not only on k and i but also on the contents of the log file that contains the dialog of the user with the NPC.

For the sake of brevity, in the following, some parts of the scoring functions will be expressed using mathematical set notation. Legend:

- L: the log file containing the sentences said by the user and the NPC. It will be regarded an ordered set of the sentences said
- L_j: The j^{th} sentence in the dialogue
- U: the subset of L consisting of the sentences said by the user
- U_j: the j^{th} sentence said by the user
- P: the subset of L consisting of the sentences said by the NPC
- P_j: the j^{th} sentence said by the NPC

Group A:

For most behavioural markers, f_{ki} is either 1 or 0 depending on the presence or absence of a certain text string (or one element of a set of strings) in the log file. p_{ki} is always 1. The latter means that for a complete fulfilment of this behavioural marker it is sufficient to produce the behaviour one time. And

[2] See te files: 'M&G - assessment by PM Game rules.xls' and 'TheProblem - assessment by PM Game rules.xls'

whatever course of actions the user takes, it cannot happen that there is no opportunity to produce the behavioural marker.

Example:

- Name of the behavioural marker: ‘Present own company’
- Belongs to skill: ‘Establishing emotional bond with customer’
- Skill is assessed in story: 'Mary and Garry'
- Behavioural marker: One should not forget to present (at least briefly) one's own company when talking to a potential project customer the first time
- Scoring rule:

  ```
  if ('Live Your Dream Inc. is in the business for 105 years.' ∋ U OR '
  Live Your Dream Inc. is specialized on outstanding weddings.' ∋ U OR '
  Live Your Dream Inc. is my company.' ∋ U) then fki = 1 else fki = 0;
  ```

- Meaning of the scoring rule: there are three possible sentences the user can say to present the company. Any of these counts as fulfilment.

Group B:

f_{ki} is 1 or 0 exactly as in group A. However, p_{ki} is not always 1 but can also become 0, depending on the course of action the user takes when playing the story. This means that the behavioural marker is not observable; the user has neither fulfilled nor failed it. In this case b_{ki} is 0.

Example:

- Name of the behavioural marker: ‘No belittlement’
- Belongs to skill: ‘Giving feedback to employee’
- Skill is assessed in story: 'The Problem'
- Behavioural marker: Major problems regarding the work of an employee should not be presented as if being harmless
- Scoring rule:

  ```
  if ('The safety is not good' ∋ U OR 'The safety is not good enough' ∋ U)
  then {fki = 1; pki = 1;} else if ('The safety is a little bit problematic'
  ∋ U) then {fki = 0; pki = 1;} else {fki = 0; pki = 0;}
  ```

- Meaning of the scoring rule: there are two possible sentences the user can say to mention the problem of the safety without making the issue appear harmless: 'The safety is not good' and 'The safety is not good enough'. Any of these counts as fulfilment. The sentence 'The safety is a little bit problematic' represents a 'belittlement', thus a failure of the behavioural marker. If none of these occurred, then there is no judgment possible on this behavioural marker.

Group C:

f_{ki} is an integer number depending on the course of action the user takes and p_{ki} is a fixed integer number specific for the behavioural marker in the respective story.

Example:

- Name of the behavioural marker: ‘Active listening’
- Belongs to skill: ‘Eliciting customer requirements’
- Skill is assessed in story: 'Mary and Garry'

- Behavioural marker: The customer should be provided with utterances such as ‘hm’ or ‘I see’, indicating that one is listening to his requirements. Remark: There are many other situations in which activelistening should be practiced. A conversation to a customer about requirements is merely the specific case of interest in this story.
- Scoring rule:

```
pki = 2;
fki = 0;
for (int j = 0, j++, j < numof(L))
   if (Lj ∋ P AND Lj == 'We would like a real Hawaiian style wedding. 100
   persons will be invited.' OR Lj == 'The decoration should give an au-
   thentic Hawaiian flavor: palm leaves on the walls and lei for all la-
   dies. Lei are traditional Hawaiian flower garlands.' OR ...
   // There are many more sentences that count as expressions of require-
   ments. For the sake of brevity they are not all listed here.
   AND Lj+1 ∋ U AND Lj+1 == 'Ok' OR Lj+1 == 'Thank you' OR Lj+1 == 'Big thank
   you' OR Lj+1 == 'I see' OR Lj+1 == 'Yes' OR Lj+1 == 'I am happy' OR Lj+1
   == 'I am satisfied') then fki ++;
if (fki >= 2) then fki = 2;
```

- Meaning of the scoring rule: Count the total number of occurrences of any of the following sentences, said by the user: 'Ok', 'Thank you', 'Big thank you', 'I see', 'Yes', 'I am happy', 'I am satisfied'. Count them however only if they come immediately after a sentence in which the NPC (the customer) said a requirement. f_{ki} is the result of this counting, but limited to maximally two. p_{ki} is in any case two. In sum, the game expects the user to do two times 'active listening' for a 100% score and one time for a 50% score.

Group D:

In the most complicated case, both the number of fulfilments and the number of opportunities depend on the course of action taken by the user. That is, both f_{ki} and p_{ki} are computed by a function that operates on the log file and returns an integer.

Example:

- Name of the behavioural marker: ‘Talk about important problems’
- Belongs to skill: ‘Giving feedback to employee’
- Skill is assessed in story: 'The Problem'
- Behavioural marker: In a feedback conversation, important problems should be talked about, at least acknowledged and not ignored or forgotten, let them come from the leader or from the employee.
- Scoring rule:

```
fki = 0;
if (∃ s ∋ U is_substring('the safety', s)) then fki ++;
if (∃ s ∋ U is_substring('Toni', s)) then fki ++;
if (∃ s ∋ U is_substring('the police', s)) then fki ++;
if (∃ s ∋ U is_substring('holidays', s)) then fki ++;
if ('Oh... I made mistakes here! It was very hectic work. I forgot many details about the
workers safety. I'm so overworked. Why don't you give me holidays?? I had no holidays
all the year. May I just take 4 days off?' ∋ P then pki = 4 esle pki = 3;
```

- Meaning of the scoring rule: The user is expected to produce communication about the four serious problems presented in the story: safety issues on a building site, a problem with the workers representative Toni, a problem with the police and the issue that the employee is

overworked and had no holidays all the year. Any communication on these counts. However, the user receives an opportunity to talk about the holidays issue only if the employee makes his complaint that he had none. Without this, the problem is not known and there is no possibility to talk about this problem. This causes the difference in the number of opportunities being 3 or 4. All other problem topics are presented right from the beginning in the form of icons and textual problem descriptions available on mouse-over.

Motivations behind design decisions

The scoring algorithm was developed with the Extreme Programming principle of simplicity (Wells, 2011) in mind: develop the simplest solution that solves the problem and add complexity as needed. This led to the following decisions:

Extract scores from the log file: while it would be possible to think of an algorithm that extracts scoring information already while playing, this would increases complexity because it always has to be monitored if there is a possibility that some information relevant for scoring arrives later.

When examining the groups A) - D), the question arises, why exactly these types of scoring functions were relevant. The answer is again simplicity and adding more complexity as needed. Concretely, the development of the scoring rules started with simple rules as in group A. Then further categories were added as needed to match the requirements imposed by behavioural markers and experiences with users. The process is illustrated by the following example: thinking about how to formulate a scoring rule for 'active listening', the group C was added as it is the property of this behavioural marker that it is useful to perform it repeatedly in a conversation. But how many times should it be done for a good performance? In preliminary user tests with the PM Game, users did not produce ‘active listening’ as often as they would in a normal oral conversation. Even users experienced in sales would stop performing ‘active listening’ after about two occurrences while they would do it for example 10 or 20 times in a similar face-to-face conversation. Thus, in the PM Game, $p_{ki} = 2$. This was the first example of a scoring rule of the group C.

Possible future extensions

The above approach aimed at an as-simple-as-possible scoring algorithm matching requirements that are discernible from properties of behavioural markers and experiences from user behaviour. As shown in Section 5.3, the scores resulting from this algorithm are valid. This, however, does not mean that there would be no potentials to improve the algorithm. Future research will show if the following possibilities result in any useful improvement of assessment quality:

- Not all observations are to be weighted equally. There can be events in social situations that ruin the entire situation and render useless all the other things that have been well done.
- There can be different levels regarding the certainty of the assessment. If one user chose a course of actions in which there occurred only one opportunity to produce a behavioural marker and another user chose a course of actions in which there were many opportunities, then the assessment results will not have the same weight of evidence.
- If a behaviour is produced but not in the right situation, this also shows some level of skill. It proves that at least the behaviour is known. (This might be a typical training effect, when people learn new behaviours but are not yet confident about when to apply them.) Thus, it might

be helpful to differentiate in the assessment results regarding breadth of the behavioural repertoire and its situational appropriateness.
- Continuing the above line of thought, the same behavioural marker could be tested in different situations. For example, for learning purposes, it could be useful to let a learner explore situations in which it would or would not be appropriate to apply a certain sales strategy.

4.8 Overall aspect: Authoring to steer stories and NPCs

Aspects of the creation of playable game stories (authoring) as well as the game functionality used to steer stories and NPCs have not been presented as a part of the method in Section 4.2. This is to keep the method open to interchangeable parts. For the assessment of human social skills, the requirement is that playing the game makes the user show behaviours relevant to the skill to be tested. Apart from this, for the assessment function, it is not important to determine what steers the behaviour of an NPC, what steers the complete sequence of a story and how to author these. However, these are important for the extensibility of the game to further skills and further stories. Extensibility is the main common requirement for all four of the use cases discussed in Chapter 2. This is the reason to present here those aspects of the simulation methods that are relevant to extensibility.

The goal of the PM Game is to assess the social behaviour of a human user. Thus, it would be plausible that the best state-of-the-art artificial intelligence, cognitive and personality modelling would be required to simulate conversation partners (NPCs). Indeed this could be an option; however, it is not necessary. In the following, a substantially simpler method will be presented that also facilitates easy authoring.

The core idea is to imitate human role-play actors. Role-play actors are professionals who play a predefined role to simulate conversation partners, in assessment centres for example. The assessment centre participant shows his or her skills in interacting with this actor. The actor needs to present the same role to all the participants in order to present equal test conditions. Thus, the role is defined in a role-play script. The participant also receives instructions. These do not define how to act, but describe the situation and the goal that should be achieved in that situation. For examples of role-play scripts, see Appendix 7.2. Authoring a story in the PM Game can be understood as writing computer-executable role-play scripts.

In the following, the most important methodical elements of authoring are presented. The focus will be on authoring the complete flow of events in the story, along with the NPC behaviour.

Remarks:

- The authoring of the elements of the social situations (Section 4.3) will not be described here. This process is straightforward and consists of editing all the texts associated with these elements, as well as choosing icons from a library or uploading new icons. For this, there are specialized editing interfaces available; see Ito (2009).
- The full details on how to author stories for the assessment of social skills are described in the guidelines in Appendix 7.7.
- The full documentation of the stories, including various courses of possible action, all contents steering the behaviour of the NPC and all general content of the game such as e.g. communication menus can be found on the accompanying CD, directory: ‘Documentation Stories’.

4.8.1 Infoobjects

The main vehicle for authoring the reactions of the NPC is to use 'infoobjects'. The name stems from 'object that brings new information into the story' and emphasizes the difference compared to other NPC answers that do not bring in new information; see the 'speech act rules' below. Essentially, infoobjects are if-then rules that specify under which conditions and NPC should behave in which way.

In the following, the conceptual idea of infoobjects, an example comparing infoobjects with traditional role-play instructions for humans, and finally the implementation in the PM Game will be shown.

Concept

An infoobject is an if-then-structure that specifies the following:

- The conditions:
  - Which sentence stub in combination with which contents to fill into the gaps of the sentence stub must be selected by the user
  - Which sentence will be produced by the game if these conditions are fulfilled
- The 'owner': An infoobject must be appended to an NPC. This means that in conversation with this NPC the infoobject is active.
- Additionally, infoobjects can have another infoobject as a predecessor. In this case, the infoobject can only be said if the predecessor has been said.

Example

The following example illustrates how an infoobject resembles traditional role-play instructions for humans. The example will start with an excerpt from a traditional role-play script for a human role-playing actor and illustrate how a very similar behavioural instruction is implemented in the PM Game to steer an NPC.

- Role-play: 'Customer Meeting' (Appendix 7.2.4)
- Context of the role-play: The user being assessed has the role of a project manager; his role-playing partner is the representative of a bank, the 'EuroBank'. They are talking about the web-platform for an e-learning project.
- Instruction for the role-player (in German):
  Beim Nachfragen oder bei Fragen wie 'sonst noch was' etc. fällt Ihnen jedoch noch ein, dass die Bildungsplattform doch noch eine Unternehmensportrait-Seite der EuroBank beinhalten sollte.
- Translation: When the assessee asks a question like 'do you have any further requirements?' you remember that you need a web page for the EuroBank.

The same is done using infoobjects in the PM Game:

- Story: Mary and Garry
- Context: The user is a manager at an event management company for exclusive weddings. Mary and Garry are the customers, which are simulated by the computer.
- Name of infoobject: 'MG on wedding2'
- Trigger: The user asks about 'the wedding'
- Predecessor infoobject: 'MG on wedding' (produces the previous piece of information as an answer to the first question about requirements. The answer is that they wish a Hawaiian style wedding)
- Text: 'We will invite all our friends and relatives… so there should be room for 100 persons.'

The resulting dialogue:

- User: ‘Could you tell me more about the wedding?’
- NPC: ‘We will invite all our friends and relatives… so there should be room for 100 persons.’

Remark: The original version of the story ‘Mary and Garry’ that was used for the validation study did not contain this particular infoobject. Thus, this dialogue will not be found in the dialogue examples in Appendix 7.4.2. Yet, it has been chosen because it perfectly illustrates the point.

Implementation

To better understand the authoring process for infoobjects in the PM Game, the following screenshots present the relevant parts of the game editors:

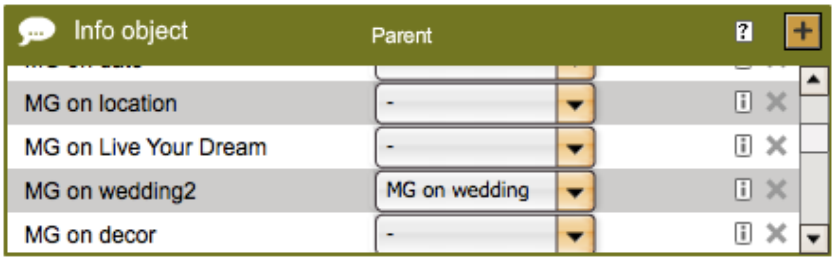


Figure 17: ‘MG on wedding’ is predecessor of ‘MG on wedding2’

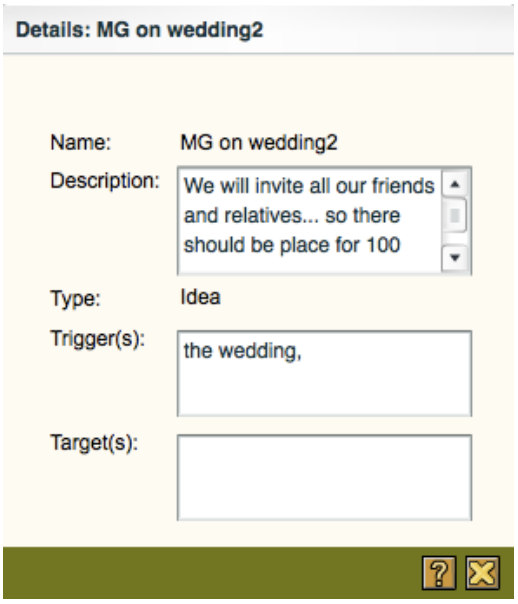


Figure 18: Text and trigger for ‘MG on wedding2’

Complete documentation

This section introduced the concept of an infoobject and illustrated it by an example. The complete documentation of all infoobjects used in the PM Game can be found on the accompanying CD, directory ‘Documentation Stories/X/Editor/Infoobjects’, where X is to be replaced by the respective story name.

4.8.2 Short stories to limit the number of infoobjects

To provide reactions for a variety of possible user sentences as the story evolves, many infoobjects must be crafted for every story. This is a laborious process. To keep the effort feasible, one of the most important guidelines is to keep the stories short. The assessment stories currently implemented in the PM Game are so short that typical log files only contain about 20 user turns. That is, the user only says about 20 different things to an NPC. See Appendix 7.4 for examples of dialogues recorded from actual game plays.

Furthermore, keeping the stories short is also a means to make sure that users actually get a chance to show or fail to show behavioural markers. If the stories were extremely long, it would be difficult to avoid some user actions that would cause the user to never reach some important parts of the story. Thus, some the potential behavioural markers for a story would never be tested. Hence, no scores for these could be computed.

4.8.3 Speech act rules

How does the game react if a user produces an unplanned sentence to which there is no infoobject available? There must always be an answer, otherwise the game stops. For this, there is a mechanism in the game called 'speech act rules'. These simply take any sentence said by the user and provide an answer to it based on the sentence stub, without considering the objects that fill in the gaps of the sentence stub. The name 'speech act rules' refers to the linguistic concept of the speech act (Searle, 1969; Austin, 1962) and can be understood in this sense: the sentence stub defines the idea of a speech act, that is, what should be done with words. A 'speech act rule' is a rule that reacts only to this idea of the speech act, but not to the 'arguments' filling in the gaps. Thus, the answers provided by such a rule are not situation specific such as 'I see', 'Yes', or 'I don't have any information on this'. To avoid meaningless communication, these answers are designed to be as widely applicable as possible. 'I see' given as the reaction to any statement by the user (e.g. 'The house is blue') perfectly realizes this. In fact, this answer can be used in any situation where the sentence stub '[person] [item] [location] is *attribute*.' requires a response. However, as another, more problematic example, 'Are you satisfied?' cannot be answered by 'I see' or 'I don't' have any information on this'. Here, 'Yes' is used as the general answer to the question. In consequence, in every story where the customer is not satisfied, this general answer must be overwritten using an infoobject.

Remark: In the validation study reported in Section 5.2, only the questions asking about information on a game object could be equipped with infoobjects (the first four questions in Figure 19).

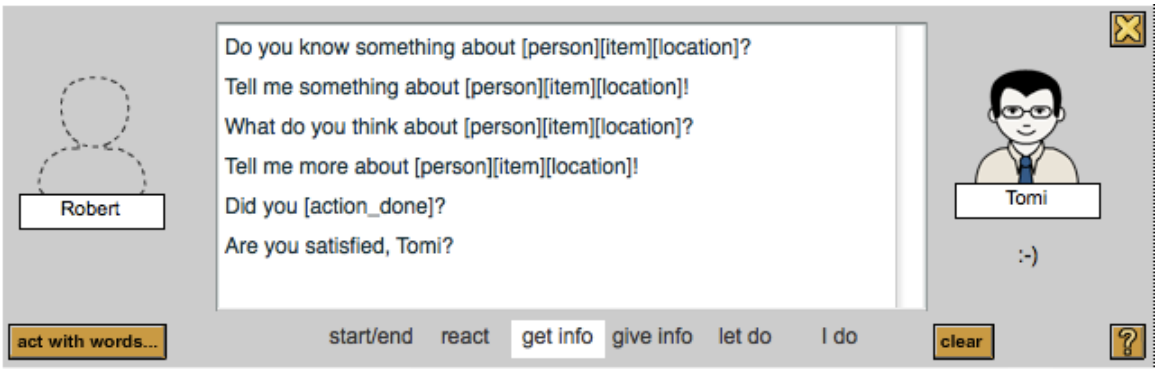


Figure 19: Sentence stubs under the menu heading 'get info'

All the other NPC answers to all the other sentences were produced using speech act rules. As a consequence, the answers to the last two sentence stubs in Figure 19 were that the NPCs were always satisfied and actions were always done. All three stories were created around these limited possibilities. For the validation study, the described limitation of the software development status did not make any difference, because the user could not distinguish how the answer 'Yes' was produced when an NPC answered the question 'Are you satisfied?' However, this should be fixed to extend the PM Game to further stories and further skills.

4.8.4 Act with words

Users are able to describe actions in their own words. This option is available via the button in the left bottom corner of Figure 19 and leads to the following window:

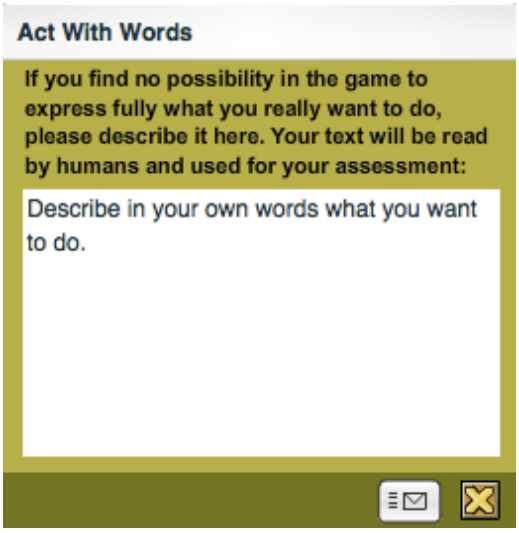


Figure 20: The interface to describe actions using the users's own words

The user instructions in the introductory story 'Hungary Hotel' teach the user to use 'act with words'. It can be used in a case where the user does not find any appropriate means to express what he or she wants to do. On one hand, this serves as a means to determine what further user actions are needed. On the other hand, reviewing the game protocols shows the plausibility that it may have had a useful function in 'venting' action intentions that could not be expressed otherwise; see the end of Appendix 7.4.2 for an example. Thus, this feature might help the user to perceive a story as a meaningful unit even if some sentences the user wants to say are lacking. That is, 'act with words' may contribute to the assessment process by avoiding frustration. On the other hand, it might also hinder the assessment quality because experience shows that some users used it too much, even where user sentences were in fact available and not difficult to find. Thus, in these cases, their action intentions escaped the assessment process because the handwritten act with word comments were not used. As a remark, the screenshot in Figure 20 above indicates that the comments would be read by humans and used for the assessment process. In the validation study presented in Section 5.2, the handwritten comments were not used for the assessment. The misleading message was given on purpose to motivate the use of act with words.

4.8.5 Means to influence the course of a story

As described in Section 4.5, users are free at any moment to say anything that can be expressed combining the communication menu and the clickable elements of the situation. Yet, there is also a way for the author to direct the flow of the events in a story to some extent, just as real events lead to consequences that change situations and action possibilities. In the PM Game, the following options are available for this.

- Each infoobject may have a list of 'targets' (see Figure 21), which are pointers to objects (NPCs, items or locations). When the infoobject is said, all the listed targets become visible at the same moment that the speech bubble containing the infoobject text appears. This is typically used to show the objects being discussed by the NPC, which allows the user to also talk about these objects. (The infoobject in Figure 21 makes Luigi talk about the holiday he wants. Simultaneous to the appearance of the speech bubble, the holiday icon appears on the game screen.) This is the most common means to add further elements to a situation.
- A further means to add, or withdraw, elements from a situation is triggering things by hidden 'goals'. However, goals are by far more general in function: they can be triggered by any attribute change in the game and thus may trigger a list of several arbitrary changes in the game situation. To not obstruct the flow of this section with numerous figures, a detailed example has been moved to Appendix 7.6.

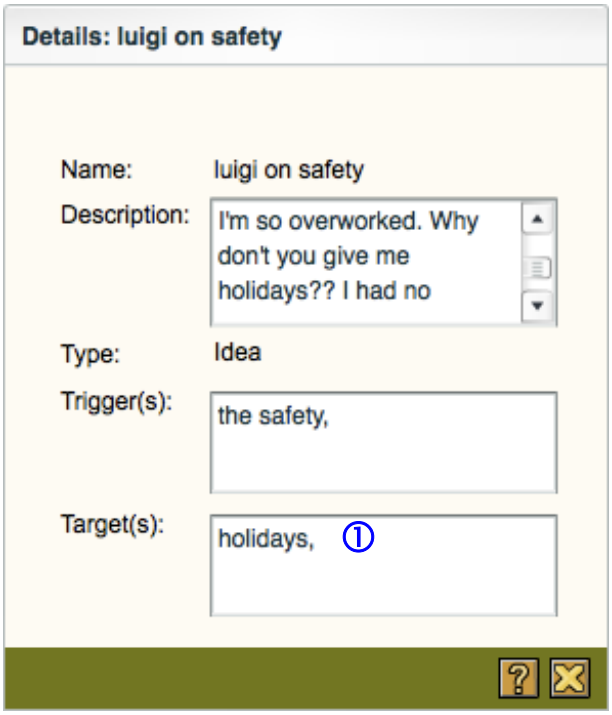


Figure 21: Infoobject with one target (1)

4.8.6 Complete example of infoobjects, speech act rules and act with words

Appendix 7.4 illustrates the concepts introduced above. It presents actual dialogues recorded while users played game stories. All the technical information has been removed from the log files, retaining only the human readable dialogues. In these dialogues, communications are indicated to stem from infoobjects, speech act rules, or act with words.

4.8.7 Summary and motivations behind design decisions

In sum, the main vehicles for creating a role-play experience for the user are infoobjects. Infoobjects are the only means used by the game to make the NPCs produce intelligent, situation specific answers. To keep the number of infoobjects limited, and hence, make the effort feasible, stories must be kept very short. Thus, the concept of authoring stories in the PM Game strongly exploits the fact that the purpose is assessment: in role-plays, for the purpose of assessment, there is a defined situation and a defined task to complete. Furthermore, because it is a test situation, the motivations for experimenting with unusual behaviour are limited. Under these conditions, the participants do not create an arbitrarily rich variety of ways to react to the situation. Experience with assessment centres shows that, after witnessing a limited number of about 10 or 20 courses of a role-play, one has seen most of what usually occurs in that role-play. Hence, because the variety of conversations is reasonably limited, the NPC reactions to all the user sentences in all the major branches of possible actions can be crafted in advance using infoobjects, as described above. The following data show that story authoring using info objects is a task of feasible size. The three stories are the ones used as a part of the validation study in Section 5.2.

Story	Number of infoobjects
Hungary Hotel (introductory)	14
Mary and Garry (assessment)	8
The Problem (assessment)	13

Table 5: Number of infoobjects in the three stories used in the validation study

This limited amount of story data has drawbacks. Obviously, with this amount of data, the game has no chance of passing a Turing test (1950), that is, making a user believe that a real human produced the answers. However, it does not need to: the goal is not the realism of the simulation but the validity of the assessment. The latter assumes only that the patterns of social behaviour the user produces are realistic. This can be summarised as 'the goal is not the realism of the simulation but the realism of user reactions to the simulation'.

In a similar vein, keeping the stories short does not come without a price. The assessment of certain skills might require longer stories. For example, one could imagine a test of behaviour under the constraint of emotional attachment evolved over time. Considering the fact that emotional involvement or even attachment does occur when reading books, watching movies or playing games in which one has a persistent character, it is not impossible to imagine that such a test could be implemented with longer PM Game stories. However, just as this would be overly long for the assessment method used in the PM Game, it would be too long for state-of-the art assessment centres, which use role-plays of a duration of a few minutes. The assessment problems are the same: expectations regarding what counts as 'good' and what as 'not good' need to be crafted in advance and with longer role-plays the number of story branches to be pre-judged increases.

Short stories are the trick to avoid artificial intelligence and cognitive modelling as well as increase authorability: stories consisting of only about 20 user actions are substantially below any normal game experience. It appears unlikely that such a limitation would be acceptable in any entertainment game. Hence, these games require the means to make NPCs react properly in a very large number of situations. To accomplish this, many state-of-the-art entertainment games are equipped with artificial intelligence. In this sense, it can be said that the use of infoobjects and short stories are the means to cir-

cumvent the need for artificial intelligence in the PM Game. The original plan was for the game to use a variety of intelligence levels and to be built on a cognitive model such as the PSI theory by Kuhl (2001). As interesting as this might have been as a research topic, it would not have helped to keep the game easily authorable: a complex cognitive architecture would steer the behaviour of the NPCs. A change would potentially cause changes in the NPC behaviour in all the stories. In such a case, all the stories would need to be validated again to assess the skills they should. Furthermore, the author of an individual story would need knowledge about what to expect from the cognitive model when crafting the story. In view of these issues, the current method of relying on infoobjects is a means of locating complexity within the stories. Thus, changes are local to the story, which is beneficial for authorability.

5 Evaluation

The PM Game is a piece of software designed to assess social skills, which has previously been done using human assessors. Thus, the most important part of the evaluation will be to measure the quality of the assessment produced by the PM Game using traditional human assessment as a benchmark. This benchmark comparison will be carried out in Section 5.4.

However, for a full evaluation, more has to be done. In the terms of design science (Section 1.3), the proposed method (Chapter 4) has to be evaluated. A first criterion for the evaluation was answered affirmatively in Section 3.1 as part of the state of the art:

- Is the function provided by the PM Game new?

According to design science criteria, a new method has further to be evaluated if it constitutes an 'addition to the knowledge base'. In determining whether this is the case or not, the following questions arise:

- Is the proposed method new?
- Does it fulfil its stated purpose of social skill assessment?

These are answered in the following sections:

- Evaluation of methodological novelty – Section 5.1
- Empirical validity study – Sections 5.2 to 5.4
- Feasibility examination of the method based on independent empirical data – Section 5.5
- Analysis of extensibility: is this only a method to assess three particular skills or is it indeed a general method for assessing many social skills? – Section 5.6

5.1 Evaluation of methodological novelty of the PM Game

Are the methods embodied in the PM Game novel or is it an application of known methods to a new application field? The purpose of this section is to investigate the state of the art from a methodological perspective. Computer-based methods will be presented that are in some or other aspect similar to what has been used in the PM Game. As a result of this section, it will be concluded what is methodically new and what not.

The main characteristics of the computer-based test (as described in Section 4.2) that will be considered are:

- Menu-based communication: the user clicks on a menu in order to communicate.
- Sentence completion: communication requires the user to complete a sentence stub.

In the following subsection, it will be examined how far these methods have been used in computer games. As the field of computer games is very large, the review will be reduced to project management and team leadership games. In the subsections thereupon, computer based methods used in any kind of software (without limitation of the application field) will be investigated to infer if similar methods have been used so far.

5.1.1 Survey of methods by game content area

Computer games designed to teach project management and team leadership were investigated by Keller (2008), a student of the present author. These address similar areas as the current contents of

the PM Game. Although many games were surveyed, only a small number were found that closely address project management or team leadership:

- Four games (SimulTrain, Viper, Simproject, and SESAM) teach project steering (milestone tracking, etc.) using a simulation approach.
- One game (Virtual Leader) teaches the user how to moderate meetings.
- One game (Infiniteams) is a multi-user game meant to teach team dynamics.

While project-management and team-leadership games represent only a small fraction of the game market and new games are constantly being developed, the results reflect some general tendencies:

- The purpose of the games is to teach and not to assess.
- There are several single-player games, which rarely involve communication. Some multi-player games involve communication, but this communication takes place between players and not between players and nonplayer characters (NPCs).

Those few games which involve communication with NPCs (of the list above, only Virtual Leader) face the problem of how the user and the NPC will communicate. The typical solution has been multiple-choice user input. In most cases, there are 2–4 predefined sentences to choose from. Virtual Leader features an advanced, diversified multiple-choice interface; see Figure 22 below.

Figure 22: A screenshot from the communication game ‘Virtual Leader’

Although the game interface is attractive and features animated three-dimensional NPCs, the actual choices available to the user are simple: one can agree with (green end) or oppose (red end) current meeting topics and choose from a few actions presented on the left side (e.g. 'finish meeting').

In conclusion, project management and team leadership is a typical serious game application area. Surveying games in this area, none featuring menu-based communication or a sentence-completion interface were found. One game was found featuring NPCs, with whom the user can communicate using an advanced multiple-choice interface.

5.1.2 Sentence completion and menu-based communication in other areas

While sentence completion and menu-based communication are rare in state-of-the-art games, these methods have been researched. In this section, these methods and associated concepts will be considered regardless of the application field.

Sentence completion and menu-based communication are overlapping concepts. One example is the sentence-completion test (SCT) by Mahlow & Hess (2004) and Hess & Mahlow (2007), in which a sentence is completed by adding sentence fragments. The fragments are offered to the user in the form of cascading menu items; see Figure 23.

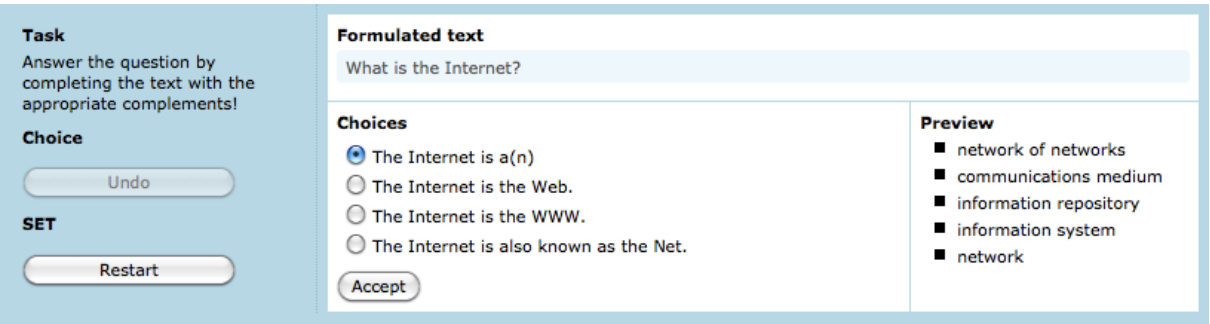


Figure 23: SCT by Hess & Mahlow (2007)

SCT are used worldwide for many purposes, such as language testing and psychological testing. For an overview of psychological tests using SCT, see e.g. Holaday et al. (2000). Many of these tests are paper-and-pencil tests or computer tests developed from previous paper-and-pencil tests.

Menu-based communication is a less widespread concept; however it also has decades of research history. An early proposal for a menu-based communication interface was the one by Tennant (1983). Its purpose was to provide an interface for a relational database. It was motivated by negative usability evaluations of the natural-language interface (NLI) that existed for that database (Tennant, 1980). Negative or false user expectations have been created when sentences formulated by a user were not understood by an NLI: users inferred that other sentences would not be understood either. The menu-based interface resolved these usability issues. The evaluation concluded that 'both experienced computer users and naïve subjects can successfully use a menu-based natural-language interface to a database to solve problems. All subjects were successfully able to solve all of their problems. Comments from subjects indicated that although the phrasing of a query might not have been exactly how the subject would have chosen to ask the question in an unconstrained, traditional [natural-language] system, the subjects were not bothered by this and could find the alternative phrasing without any difficulty.' (Tennant, 1980, p. 155).

This problem of NLIs, which is resolved by menu-based interface, is termed the 'habitability problem': an NLI is considered habitable if users can express everything needed to complete a task using a language they would expect the system to understand (Davis, 2009). In the context of the assessment of social skills, the relevance of the habitability problem is apparent: criteria to assess social skills using observed communications generally address the semantics (e.g. the behavioural markers currently used in the PM Game, see the Sections 4.6 and 4.7). However, computer understanding of natural language is currently beyond the state of the art. Thus, judgment of social skills is closely associated with habitability: to be able to assess social skills, an interface needs to make the user produce only (or mostly) communications the computer understands. Thus, while it is hard to imagine a formal proof for this, it appears reasonable to assume that the communication interface of a computer software meant to assess of social skills should be habitable.

A further point relevant for social-skills assessment is Tennant's observation cited above: 'Comments from subjects indicated that although the phrasing of a query might not have been exactly how the subject would have chosen to ask the question in an unconstrained, traditional [natural-language] system, the subjects were not bothered by this and could find the alternative phrasing without any difficulty.' A similar observation has been made regarding the PM Game: in many cases, the sentence stubs available did not permit users to say what they wanted in exactly the way that they would have formulated it. However, in most cases they could express what they wanted.

The above idea of a menu-based natural-language interface did not find such widespread application as might seem likely based on its advantages in terms of usability. However, menu-based communication has garnered some interest recently both as a means of safeguarding children and as an end-user interface to the semantic web.

Probably the best-known example of menu-based communication to safeguard children from inappropriate communications is the 'SpeedChat' function in the massively multiplayer online game ToonTown (see e.g. Mike & Warner, 2005).

Figure 24: SpeedChat in the game ToonTown – an example of menu-based communication

Another application field for menu-based communication (or guided communication in general) has been brought about by the recent interest in providing semantic-web user interfaces for casual end-users (see for example Kaufmann, 2009). In a paper entitled 'Talk to Your Semantic Web', Tennant proposed a newer version of his previous research on menu-based communication for this purpose (Thompson, Pazandak & Tennant, 2005). See Figure 25 below.

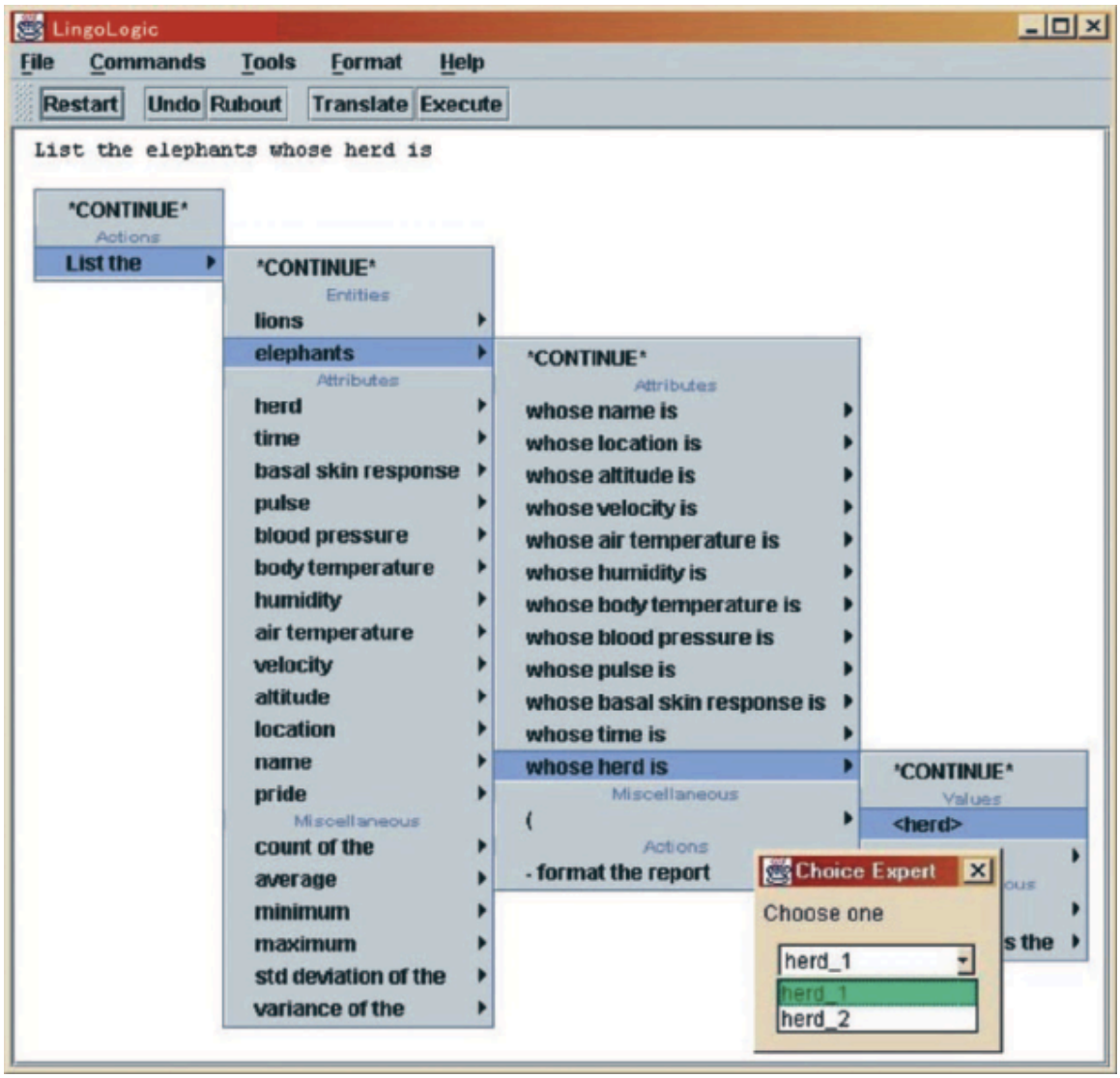


Figure 25: LingoLogic – an example of a menu-based interface to access the semantic web

To summarise this section, menu-based communication is sometimes used for specific purposes, such as to safeguard children or to solve the habitability problem when accessing a database on the semantic web. In all the menu-based communication solutions cited above, there is no possibility to combine elements of the situation with sentences from the menu. The most important insight from this section is that habitability is arguably a requirement for the interface of any computer software that is to assess social skills.

5.1.3 The simulation most similar to the PM Game

Stevens (1989) developed an e-learning system to train users to conduct code inspection. It taught theory and provided a practical simulation in which the user could try out his or her knowledge. This simulation implemented several of the ideas used in the social-skills assessment method proposed in this thesis. The interface with which the user could form sentences had the same core idea: it combined menu-based communication with sentence completion in order to talk to NPCs. The main communication menu consisted of the following items:

Clarity
Correctness
Simulation
Consistency

This menu corresponds in its function to the main menu of the PM Game:

start/end react get info give info let do I do

To each of the menu entries in Steven's e-learning solution, there was a submenu of sentence stubs similar to those in the PM Game:

<object> is (not) an ERROR. (? !)

<object> has (in)CORRECT <programming construct-aspect>

<object> does (not) FULFILL the REQUIREMENTS.

<object> when <code-action> causes an ERROR.

<object> is (un)NECESSARY.

<object> is in the WRONG POSITION.

<Object> SHOULD BE <(user input ?), variable-value-boundary>

Figure 26: Example of a submenu

The following screenshot illustrates the interface as a whole:

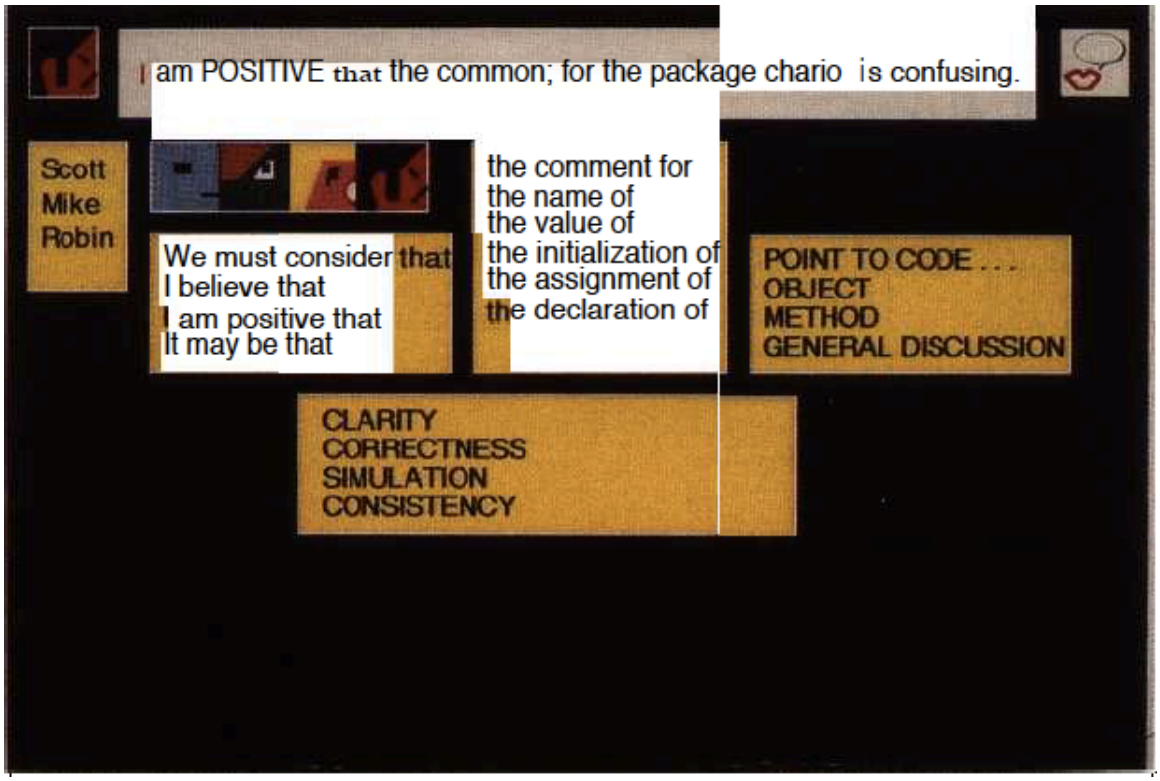


Figure 27: The user interface most similar to the PM Game

There are not only similarities but also differences to the PM Game and to the method proposed in this thesis. The above simulation features no assessment of social skills and no modelling of elements of social situations, which could be used in the sentences (Part 1a of the method; see Section 4.3). The social situation is presented to the user in the form of video fragments showing actors who talk and respond to what the user has said. However, the user cannot incorporate objects, ideas, or issues from the video into his or her communications. One can talk about precrafted problems in the code.

This simulation approach has been lauded for its potential uses in education. Dede (1995, p. 49) says: 'While the application built by this project focused on code inspection as the skill to be trained, through similar means preparation could be provided for a wide variety of work-related situations that involve social interaction within a limited range of formalized behaviors (e.g., learning to be a customs agent, developing skills in job interviewing). The educational effectiveness of this application was assessed both by the Southwest Research Institute and via a doctoral thesis at Carnegie Mellon University (Christel, 1994). The results of these evaluations document that this simulation is both instructionally effective and highly motivating for participants'.

Why is this simulation, which has been evaluated so positively, not used worldwide in universities and companies teaching code inspection? Many reasons can be imagined, for instance licensing, but one reason is certain: the learning content is outdated, as it referred to Ada program code, which is not commonly used any more. Thus, all the video recordings would have to be remade. In a similar vein, why do our educational institutions not give instruction in the application fields foreseen above using this simulation? Considering the potential applications, the current generation of learners could all be trained with this simulation, using it, as Dede says, in a wide variety of work-related situations that involve social interaction within a limited range of formalized behaviors (e.g., learning to be a customs agent, developing skills in job interviewing)'. There is no discussion in the article of how to extend the system to similar stories or to learning other skills. Judging from the description, it

definitively would have required new video recordings of actors and probably substantial programming effort as well. Experts in the learning area would have no chance to accomplish all this without advanced IT skills or personnel, filming equipment and actors.

In short, a central criterion of successful e-learning solutions is content authorability (changeability and extensibility), and this criterion was not fulfilled in this case. The PM Game differs from Stevens' simulation approach with regard to this criterion, because one of its major features is its radical simplicity – featuring no advanced graphical simulations. In addition to the feature of assessment, this can be considered the most important difference between the PM Game and the e-learning approach of Stevens (1989) discussed in this section.

5.1.4 Summary

Several computer based methods have been presented that have not been used to assess social skills but bear similarities to the PM Game in their methodical parts. No method could be found that relies on the combination of core methodical parts by which the PM Game brings about assessment of social skills.

5.2 Validation study: method

5.2.1 Study overview

To validate the PM Game, a comparison study was carried out. Participants had three skills assessed by playing the PM Game and had the same three skills tested by a well-established method of skill assessment (audio recordings of a conventional dyadic role-play task scored by human assessors).

Tested skill	Test situation	PM Game story	Role-play
- Giving feedback to employee	An employee performed a larger task; the performance was mostly good, but there are issues to be resolved.	'The Problem'	'Feedback Talk'
- Establishing an emotional bond - Eliciting requirements	A first meeting with a new customer interested in a complex service for which a longer customer relationship is required.	'Mary and Garry'	'Customer Meeting'

Table 6: Role-plays and PM Game stories used in the validation study

For simplicity, skills ('the skill to carry out a professional feedback conversation with an employee') will be described by the task ('giving feedback to employee'). Uniting tasks and skills follows the concept of task-based assessment centres, as opposed to the dimension based assessment centre approach. In the latter, skills are constructs such as 'argumentation strength' scored across different tasks (e.g. Lance, 2008).

The following sections describe why this method of validation has been chosen and how it has been carried out.

5.2.2 Reasons for the choice of the validation method

Reasons for task-based skill assessment

Whether task based assessment centres or dimension based assessment centres are better is an ongoing debate (Lance, 2008). Dimension based assessment centres have a broader research basis from decades of assessment psychology research. The newer concept of task-based assessment centres has, however, an advantage that is particularly important for the comparison study outlined in Section 5.2.1 above: it has better discriminative validity (Lance, 2008). That is, skill scores reflect more exactly the skills they are supposed to measure and there is less correlation among scores of different skills. Hence, using the task based assessment centre approach makes the study more sharp in either showing no correlation or showing a high correlation between the human and computer scores.

Reasons for role-playing

Role-play tasks scored by humans have been chosen because they are accepted by peers as representing social skills (e.g. Funke & Schuler, 1998). Furthermore, they are widely used in assessment centres for educational and job selection purposes.

Reasons for the contents of the role-plays and game stories

As shown in Table 6, three skills have been tested both by the PM Game and by the role-plays. To emphasize the practical relevance of the game, three business skills have been chosen. To demonstrate the breadth of the game's assessment possibilities, the following choices were made:

- Different kinds of conversation partners were included (customer and employee)
- Different kinds of skills were selected: the rather 'task oriented' elicitation of requirements and the more 'person oriented' establishment of an emotional bond as well as giving feedback, in which these two aspects may conflict.
- To receive a first impression of whether the game can distinguish among different skills apparent in the same situation, one PM Game story and one role-play have been conceptualized to test two different skills.

5.2.3 Subjects

Participants were recruited by a psychologist. The author provided the following guidelines:

1. All persons must be able to read English fluently and quickly (the PM Game was in English).
2. All persons must be able to speak and understand German fluently (the role-playing was in German).
3. Both sexes should be distributed as equally as possible.
4. Participants should have mixed fields of study and occupation.
5. No IT students or professionals (their advantage in computer handling could create noise variance).
6. No medical students or professionals. At Zurich University, medical students were often excluded from studies targeting leadership and social behaviour because of the perception that medicine's leadership culture might be more hierarchical and that their performance could not be measured using the same criteria; for example, they may have different standards regarding giving proper feedback.
7. Persons should be young professionals or students. The younger generation can be assumed to have grown up with computers, have a more natural attitude towards them, and a general familiarity with computer games. Specific computer skills were tentatively not used as a criteri-

on, because the computer-based assessment should apply to the general population. However, whether someone has grown up with computers is a significant difference.

8. Participants must have no previous experience with the PM Game.

There were 47 participants, ranging in age from 19 to 39 years (M = 27.17, SD = 4.6). Two participants were excluded because they were outliers concerning age (see Point 7 from the guidelines above): both were 45 years old. While they were permitted to complete the assessment centre, their MP3 recordings from the role-plays were used for assessor training purposes. Participants were 59.2% female. Most participants had professional experience, varying highly in duration and kind. A total of 59.6% reported at least minimal experience in leadership. Most participants were students; some participants had already graduated or had not pursued a degree. The students all came from Swiss universities, mostly from the University of Zurich. The majority studied psychology, journalism, or sociology. See the accompanying CD for full details, file name: 'Raw data participants.sav'. Participants received a small compensation of CHF 20.

5.2.4 Study procedure

As a first step, several informal pretests were carried out using the PM Game in order to gain experience with usability, determine necessary improvements, find assessable skills, and test the assessment criteria.

In the validation study, the following procedure was carried out:

An assessment centre was run by a psychologist with assessment centre training and experience. The author provided him the role-play documents (Appendix 7.2) and administrative access to the PM Game. Participants were invited by the psychologist. In the assessment centre, the author was not present, to avoid any risk of influence. The assessment centre executed the following procedure for each participant:

1. A short introduction including a statement of data confidentiality
2. A questionnaire on general participant data
3. Two dyadic role-plays in which the psychologist served as the role-playing partner. The psychologist had training and experience in assessment centre role-play acting.
   3.1. 'Feedback Talk'
   3.2. 'Customer Meeting'
   The role-plays were voice-recorded as MP3 files and labelled with the participant's code. Each participant had a random city name as code.
4. The participant played the PM Game. Each game story represented a role-play in which a non-player character served as the role-playing partner. The sequence was:
   4.0. Short video tutorial on game handling
   4.1. 'Hungary Hotel' (training story to exercise game handling)
   4.2. 'The Problem'
   4.3. 'Mary and Garry'
   Playing the PM Game resulted in log files saved under an index number linked to the participant's real name in the PM Game database.
5. Further questionnaire

6. Payment of the CHF 20 compensation to the participant. Neither feedback nor a score was provided to the participant at that time in order to avoid the possibility of influencing other participants.

The order of Steps 3 and 4 to 5 was randomized to control for possible memory effects (see Section 5.3.4).

5.2.5 Development of test contents

The contents of the role-play tasks are described in Section 7.2. For the contents of the PM Game tasks, see Section 7.1. This section is to document briefly the development of these contents as part of the preparations for this validation study.

The original source of the test contents are a role-play in a lecture by Prof. Dr. Martin Glinz held at Zurich University (eliciting customer requirements) and several years of industrial project management experience of the author in the case of 'establishing an emotional bond in a first customer meeting' and 'giving feedback to employees'. Drawing from these, role-plays with similar contents were developed, tested and used with several hundred participants (students, university employees and company employees) in a lecture (Stoyan, 2008). The present form of the role-plays (as used in this study) was developed by Ebert (2008) in order to evaluate the learning effect of the project management course of Stoyan (2008). The PM Game stories were developed by the author of this thesis to test the same contents as Ebert's role-play, as close as this is possible with the PM Game. The scoring criteria used by Ebert were refined by the author to yield the behavioural markers used in this study. The behavioural markers were validated by two psychologists. The exact scoring rules (expectation levels and conditions, see the accompanying CD, files: 'M&G - assessment by PM Game rules.xls', 'TheProblem - assessment by PM Game rules.xls') used to obtain the game scores for the study participants were developed by the author using preliminary user tests. The method for these tests is described in Appendix 7.7.

5.2.6 Scoring of the MP3 recordings

Two trained human assessors assessed the MP3 recordings using the behavioural checklist method (e.g. Hennessy et al., 1998). The full behavioural checklists used for each of the skills can be found on the accompanying CD, file names: 'M&G - assessment by humans' and 'TheProblem - assessment by humans.xls'. The assessors were the author (with four years of project management experience, a book on project management, four years of experience in teaching team leadership and project management, and assessment centre training) and a psychologist (with assessor training and experience, training in team leadership and project management). This corresponds to the common procedure of using both psychologists and managers as assessors; compare, for example, Brummel et al. (2009).

The assessment procedure for each participant was the following:

1. Each assessor performed the following steps alone:
   - Listened to the MP3 recording as many times as needed
   - Noted observations in the checklist
   - Where required by the checklist, formed a mark (1, 2, 3, 4, or 5. 5 was the best mark) for groups of observations (sub-skill marks).
   - Gave an overall mark for the skill where required by the checklist

2. The two assessors then examined each other's observations and marks, and did the following:

- If observations differed, they were discussed, and if needed, the assessors listened again to the MP3 recordings
- In most cases, the checklist contained guidelines on how to form the marks from the observations. If one assessor believed that the other had not correctly followed these guidelines, this was discussed. These discussions resulted in either detection of unintended deviations from the guidelines (i.e. errors in scoring) or acknowledgement that a special case provided a good reason to deviate from the guidelines
- No other discussion took place on the individual marks given by each assessor
- Finally, they agreed on a common overall mark for the skill

The MP3 data on the two excluded participants were used for test-runs of the above MP3 scoring procedure. These test-runs served both to train the assessors and to help them improve the formulation of the checklist items.

5.2.7 Scoring of the log files

The log files have been scored according to predefined rules, see Section 4.7. As the scoring module of the PM Game had not yet been implemented, the rules were executed by hand, reading the log files of each user and writing the results into an Excel table (see the accompanying CD, files: 'TheProblem - assessment by PM Game rules.xls' and 'M&G - assessment by PM Game rules.xls'). As this makes no difference to the study, the results of the log files scoring will be referred to as 'computer assessment'.

5.3 Validation study: results

5.3.1 Main result

For all assessed skills, there was a significant positive correlation between the participant´s role-play scores assessed by humans ('human assessment') and their game play scores assessed by rules ('computer assessment').

Skill	Spearman's rho of computer vs. human assessment	N	Number of behavioural markers used for the assessment
Giving feedback to employee	.701*** (.000)	47	Human: 24 Computer: 8
Establishing emotional bond with customer	.511*** (.000)	46	Human: 13 Computer: 7
Eliciting customer requirements	.383** (.009)	46	Human: 5 Computer: 3

Table 7: Correlation of computer assesment versus human assessment

Legend and remarks on the columns of Table 7:

- Column 'Spearman's rho':
  - In this thesis, the following notation is used to quote correlations:
    correlation coefficient (significance)
  - The Spearman correlation is used because the data are (strictly speaking) ordinal scaled; there is no guarantee that one unit of difference on the scale always signifies

the same difference in abilities. Many researchers argue that the units of ability scales are usually quite evenly distributed and that the ability scales are somewhere between the ordinal and interval scales. Theories and testing conditions exist that justify statistical computations on ability scales as interval scales (Rasch, 1961). In this dissertation, however, there is no need for interval scaled scores, as its argument relies on correlations, which work on ordinal scales as well, using Spearman's rho.

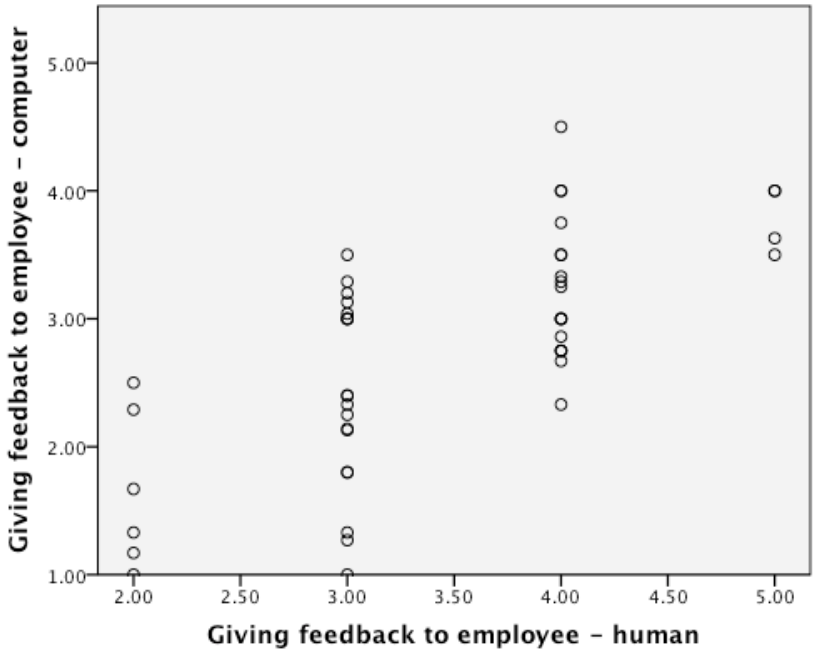
- Column 'N': One participant did not play the game story 'Marry and Gary', which tests the skills 'establishing emotional bond with customer' and 'eliciting customer requirements'. The participant just said 'hello' and left, hence the difference of N = 47 vs. 46.
- Column 'Number of behavioural markers': For a better understanding of the different correlation numbers, the number of behavioural markers is presented in Table 7. This concept was introduced in Section 4.6 and describes how many pieces of behavioural information have been used to assess the performance of each participant. As Table 7 indicates, for the considered three skills, the correlation substantially increases with the number of behavioural markers. This raises the hope of a potential to achieve a high fit between human and computer scores by increasing the number of behavioural markers.

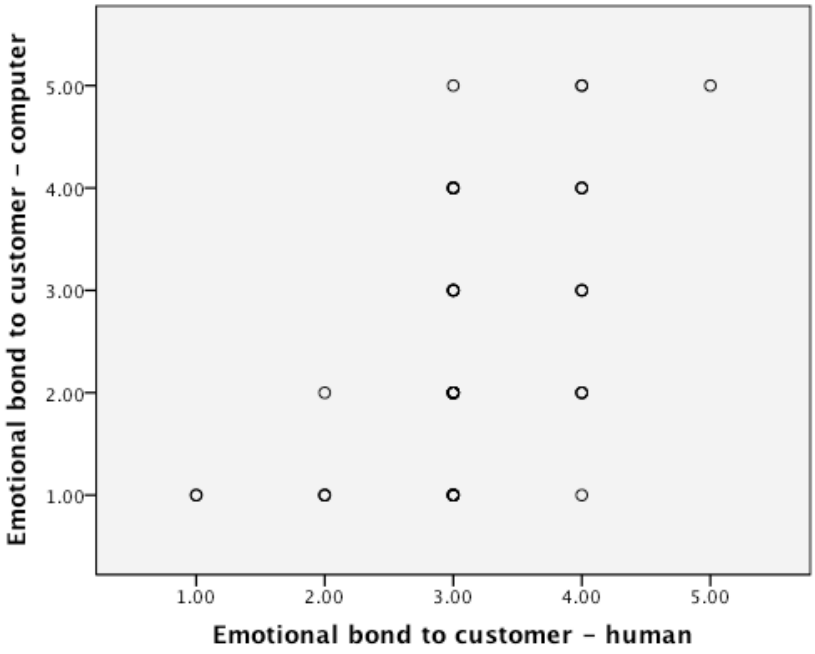
The correlations presented in Table 7 above are part of the reason for the high-profile claim of having developed the first computer-based assessment method for social skills. Thus, particular rigour is appropriate in examining if this claim can be substantiated. This will be done through the following steps:

1. Verify the above dependence between computer assessments and human assessments (Section 5.3.2 to 5.3.5, and Section 5.3.8).
2. The above dependence qualifies only the shared variance but does not imply a statement on the rest of the variance. Does the latter also stem from aspects of the social skills? This question is particularly critical to answer in any validation process, as it is, by definition, not resolved with a simple correlation. Hence, the following examinations are provided:
    - Exclusion of other possible source of variance (Section 5.3.7)
    - As far as available, the identification of statistical variables that should correlate with the computer scores if they are a valid measure of the assessed social skills (Section 5.3.6)
3. After having examined and verified the empirical evidence of dependency, questions arise: How good are these results? How large should such correlations be? Are these expectations fully met? These questions will be answered through the literature comparison in Section 5.4.

5.3.2 General statistical robustness

Though Spearman's rho is a robust measure of correlation (Croux and Denom, 2010), for additional certainty that the correlations are not caused by several outliers, scatter plots will be presented. They indicate a linear relationship between the human and computer scores for all three skills.


Giving feedback to employee – computer
5.00
4.00
3.00
2.00
1.00
2.00
2.50
3.00
3.50
4.00
4.50
5.00
Giving feedback to employee – human


Emotional bond to customer – computer
5.00
4.00
3.00
2.00
1.00
1.00
2.00
3.00
4.00
5.00
Emotional bond to customer – human

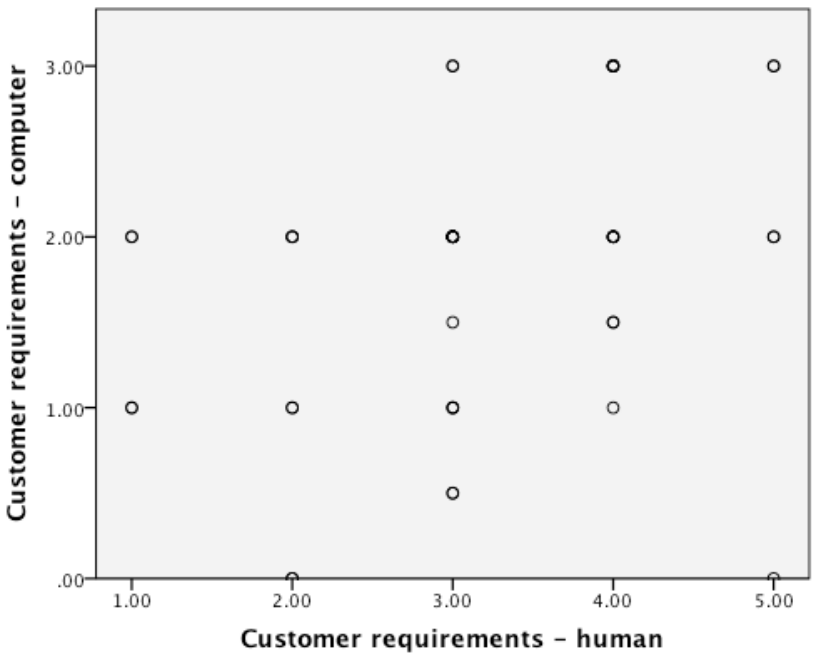


Figure 28: Scatter plots for computer versus human assessments

5.3.3 Robustness to potentially influential scoring decisions

Examining the human and computer scoring methods, are there any decisions that might have had a large influence on the result?

Giving feedback to employee

For the skill 'giving feedback to employee', the following procedure was carried out to form the human scores: 24 behavioural markers were observed for each participant in this role-play task. These observations were noted in 'yes'/'no' form. These 24 behavioural markers were grouped into 8 sub-skills. Each received from each assessor an integer mark from 1 to 5. From each of these sub-skill marks, each assessor formed an overall mark for the skill 'giving feedback to employee'. Finally, the assessors consulted, using the procedure described in Section 5.2.4, and agreed on an overall mark for the participant's performance in the skill 'giving feedback to employee'.

Altogether, to obtain a single human skill score for one participant, the following numbers of decisions were made as part of the scoring procedure:

Decision:	Number of decisions
Observe behavioural markers	2 x 24
Form sub-skill marks	2 x 8
Discuss and decide on an overall mark	1

Table 8: Human assessment decisions in judging the skill 'giving feedback to employee'

Obviously, the last step in forming the overall mark represents a threat to the robustness of the scoring method. This step is done only once per participant and has a substantial effect on the correlation between the human and computer scores. Regarding the computer assessments, the algorithm for 'giving feedback to employee' is the simple sum of 8 computer assessments. Thus, in the computer assessment, no single point substantially influences the scores.

As a test for robustness, the correlations of the computer assessments and the human assessments will be calculated again, using the computer assessment in its original form, but modifying the human assessment. The last step in forming the overall marks by humans will be replaced by the average of the sub-skill marks. This avoids individually influential scoring decisions and is thus a test of robustness. It is to be noted that this change is not without influence on the quality of scoring; it will destroy the assessment information the human assessors produced by weighing the sub-skill marks, as it is appropriate for an individual case, as, for example, if a participant did not permit the recipient to express a single word of his or her own opinion. Such rude behaviour may colour the whole impression of the role-play, and the human assessor may decide to give more weight to this observation. Such human insights are destroyed when replacing the overall mark agreed on by the two assessors with the average, thus giving all parts of the scoring sheet equal weight. On the other hand, the average is numerically more exact, as it allows more than just the integer marks (avoids the rounding error).

Resulting correlations:

Skill: giving feedback to employee	Spearman's rho
Average of human sub-skill marks versus computer score	.582*** (.000)
For comparison with the original values: human score vs. computer score	.701*** (.000)

Table 9: Correlations for 'giving feedback to employee' whithout influential human decisions

Establishing emotional bond with customer

For this skill, the situation is similar:

Decision:	Number of decisions
Observe behavioural markers	2 x 13
Form sub-skill marks	2 x 3
Discuss and decide for one overall mark	1

Table 10: Human assessment decisions to judge the skill 'establishing emotional bond'

For the computer assessment, the scores from 7 behavioural markers have been added up to obtain the total score for the participant.

Again, as a test for robustness, the correlations of the computer and human assessments will be recalculated using the computer assessment in its original form and replacing the overall marks with the average of the sub-skill marks.

Resulting correlations:

Skill: Establishing emotional bond with customer	Spearman's rho
Average of human sub-skill marks versus computer score	.429** (.003)
For comparison with the original values: human score vs. computer score	.511*** (.000)

Table 11: Correlations for 'establishing emotional bond' without influential human decisions

Eliciting customer requirements

Similarly, the situation for this skill is:

Decision:	Number of decisions
Observe behavioural markers	2 x 5
Form sub-skill marks	2 x 4
Discuss and decide for one overall mark	1

Table 12: Human assessment decisions in judging the skill 'eliciting customer requirements'

For the computer assessment, the scores from three behavioural markers have been added up to obtain the total score for the participant.

Again, as a test for robustness, the correlations of the computer and human assessments will be recalculated using the computer assessment in its original form and replacing the overall marks with the average of the sub-skill marks.

Resulting correlations:

Skill: Eliciting customer requirements	Spearman's rho
Average of human sub-skill marks vs. computer score	.440** (.002)
For comparison the original values: human score vs. computer score	.383** (.009)

Table 13: Correlations for 'eliciting customer requirements' without influential human decisions

5.3.4 Excluding memory effects

In this validation study, the PM Game and role-plays assessed the same skills; thus, the test contents cannot be made to be completely unrelated. Hence, the risk that the participants' performances in the PM Game and the role-plays are related because participants have repeated some of their behaviours.

To diminish this risk, the situations presented in the role-plays and the PM Game differed in regard of the following aspects:

- Person and company names
- Branches of business
- Project contents
- Languages (English in the PM Game versus German in the role-plays)

In spite of these measures, there is still a risk that some memory effects occurred and that the correlations were partially due to this. Thus, to control the memory effects, a randomization strategy was carried out: approximately half the participants did both role-plays first; the rest did both PM Game stories first (see Section 5.2.4).

In order to evaluate if a memory effect occurred, the mean scores of the participant groups have been computed. With a memory effect, the scores of the test taken second could have been altered. Were this the case, the sign of the changes (increase or decrease) of the mean scores would be the same for a given type of test (i.e. computer scored or human scored).

	Computer: Feedback	Human: Feedback	Computer: Emotional	Human: Emotional	Computer: Require-ments	Human: Require-ments
PM Game first						
Mean	2.6448	3.4000	2.5200	3.1600	1.9000	3.2800
N	25	25	25	25	25	25
Std. dev.	.92087	.81650	1.19443	.55377	.77728	1.20830
Role-play first						
Mean	2.8732	3.4545	2.4762	3.1364	1.6667	3.0000
N	22	22	21	22	21	22
Std. dev.	.81833	.85786	1.47034	.99021	.92646	.97590
p (eq.v.)	.376	.824	.912	.919	.358	.391
P (n.eq.v.)	.373	.825	.913	.922	.366	.385

Table 14: Mean scores of the two participant groups

Legend:

- Feedback: Giving feedback to employee
- Emotional bond: Establishing emotional bond with customer
- Requirements: Eliciting customer requirements
- p (eq.v.): significance (two-tailed) in case of equal variances assumed.
- p (n.eq.v.): significance (two-tailed) in case of equal variances not assumed
- The arrows point from the mean scores of the test taken first to the mean scores of the test taken second: they point from a not influenced test to a test possibly influenced by a memory effect. The plus/minus sign at the arrow tips indicate an increase or decrease in the mean scores.

Table 14 shows that the differences are small (do not pass the t-test) and that there is no consistency in the directions of the possible influence. An indication would be present if all arrows pointing upward or all arrows pointing downward had the same sign. Both cases have inconsistencies, however. In conclusion, there is no indication of a memory effect. It should be noted that, as the variables are not interval scaled, the computation of means is not permissible. Thus, the above data can be of only approximate value.

In order to further elucidate if correlations were caused by memory effects, groupwise correlations have been computed. This is a means of detecting one-directional memory effects—situations where one type of test (computer scored or human scored) significantly influenced the other. If a one-directional memory effect has occurred, one of the two participant groups should have larger correlations. It is not a means of detecting memory effects, where both tests influence each other to a similar extent, because that would not cause differences in correlations.

Spearman's rho for computer versus human scores	Giving feedback to employee	Establishing emotional bond with customer	Eliciting customer requirements
PM game first	.691*** (.000) 25	.383 (.059) 25	.337 (.100) 25
Role-play first	.721*** (.000) 22	.619** (.003) 21	.315 (.164) 21

Table 15: Correlations of computer vs. human scores for the two participant groups

Legend: The entries in the table fields are the following: correlation coefficient, (significance), N

In Table 15, no participant group had consistently better correlations. Thus, there is no indication of a memory effect.

5.3.5 Excluding undesired human influence

The following measures have been taken to eliminate any chance of undesired human influence on the results:

- For the MP3 recordings of the role-plays, an assessment method has been chosen that is almost independent of the person of the assessor.
- A rigorous blinding scheme has been carried out.
- Most importantly, the full details of all assessor and rule-based judgments as well as all game play logs and MP3 recordings are provided on the attached CD and can be checked by anyone.

Assessment method

The MP3 recordings were assessed using the behavioural checklist method. This method has the best performance in regard of being assessor-independent (e.g. Hennessy et al., 1998). This method presents the assessors with a list of predefined behaviours and asks for a judgment about whether the behaviour is present (or to what extent). In a typical example, the assessor might have to judge if an argument for a quoted price was sound (or to what extent).

This study made the judgments even more independent of the assessors by subdividing the checklist into small behavioural items. A typical example is whether the participant presented the name of his own company or not. This would not be feasible for any industrial assessment centre, as it makes the checklist substantially longer. A longer checklist increases the number of times the recordings have to be heard to fill up the complete checklist with observations. However, for the research purposes of the current study, this additional effort was undertaken in order to receive highly exact assessments and make the judgments independent from the person of the assessors. For example, anyone listening carefully to the MP3 would make the same observation about whether a company name that had to be mentioned was mentioned or not. Assessor influence was possible on only a few checklist items. One example is ‘active listening’: showing the conversation partner that one is carefully listening by saying ‘hm’, ‘yes’, ‘I see’ or ‘interesting’. The decision about whether the assessee produced sufficient ‘active listening’ when talking to the customer was up to the personal judgment of the assessor, as little about that item can be formalized.

In this study, not only the smallest judgments but also their aggregation to sub-skill marks left little room for assessor influence: the weight of a single observation’s contribution to the rating of the sub-skill was determined by the checklist. Furthermore, the checklist provided the authors with the average of the sub-skill marks as a guide in determining the overall rating for the whole skill. Assessors had the freedom to override these checklist rules. However, they felt the need for this only in single cases during the assessment procedure. In addition, the reasons for this were noted in the assessment sheet.

The assessment procedure forbade any discussion of assessor ratings. It permitted discussion only of the following:

- observations
- possible errors in following the checklist

- determining the common overall rating.

See Appendix 7.3 for a signed document about the assessment principles.

Blinding scheme

While assessing the MP3 recordings, the assessors were blind to the PM Game performance of the participants. While executing the rules to compute the PM Game scores, the author was blind to the performance of the role-play participants. This was achieved by indexing the MP3 recording (and all survey data from the assessment centre) by participant code. Participant codes used city names, such as 'Alamo'. The log files—the only source of information about the performance of the PM Game participants—were indexed according to the participants' real names (on the accompanying CD, the city names have been inserted).

5.3.6 Confirmation of dependence using related variables

If a variable represents a skill measure, it should exhibit a statistical relationship with the variables representing past opportunities to gain that skill or other judgments related to that skill. In this section, such 'related variables' will be investigated.

When individuals are confronted by social situations and their perceptions of others' reactions, this provides opportunities to learn social skills. Hence, as an indication of the game's valid social skill assessment, the scores can be examined for how well they correlate with the self-reports on experience regarding the same social skills.

This approach has been taken for the skill of 'giving feedback to an employee'. Confirming the expectation, there is indeed a significant correlation in the computer scores for the self-reports on leadership skills and a somewhat weaker non-significant correlation in the human assessments.

	Spearman's rho	N
Computer assessment of giving feedback to employee vs. self-judgment of leadership experience	.350* (.016)	47
Human assessment of giving feedback to employee vs. self-judgment of leadership experience	.256 (.082)	47

Table 16: Correlations between leadership experience and 'giving feedback to employee'

For the other two skills, 'establishing emotional bond with customer' and 'eliciting customer requirements', a different approach was taken because the above approach relies on the assumption that study participants understand exactly what kind of experience is asked for. The situation presented in 'Mary and Garry' involves a customer conversation about a complex service with many parts that involves trust and building up a customer relationship that will continue for months. This situation is substantially different in its challenges from many other customer conversations such as for example an informal chat at a cashier's desk. Thus there had to be an exact explanation specifying the required experience. Such an exact explanation in turn could have influenced participants' performance by providing clues regarding the assessment criteria. An exact explanation seemed, however, inevitable as participants were expected to be rather inexperienced in customer conversations. Looking at the data, this was justified as 26 of 47 participants reported 'zero' as their number of customer conversations; see the accompanying CD, file: 'Raw data participants.xls'.

In view of the above, in order to create a related variable for the game story ‘customer conversation’, participants were simply been asked for the number of customer conversations they had had so far. In order to avoid lengthy definitions of what kind of customer conversations are meant, participants first played the game and then answered the objective question of how many customer conversations they had had. This was done in the hope that playing the game would prime participants to the type of customer conversation meant, without requiring any further explanation (and asking an objective question avoids the risk of influencing the skill self-judgments through impressions drawn from taking the test). Unfortunately, this strategy did not lead to usable results. The number of customer conversations exhibited no relationship to either the computer or the human assessments. An indication against the validity of the game would have occurred if the established means of skill assessment (human assessment) had correlated and the computer assessment had not. An indication in favour would have occurred if both had correlated. Neither was the case, hence there was no gain of information.

Finally, among all surveyed participant variables, participant age could be indirectly related to various social skills: a higher age means more past opportunities to learn skills, which may lead to better skills. A computation of correlations and an examination of scatter plots show that none of the computer scores for any of the three skills exhibited any relationship to age. This is easily explained, however, by the fact that the human assessments of the role-plays as well as self-judgments of leadership skill also did not show any relationships to participant age. Hence, it can be concluded that, in this particular sample of 47 participants, the older ones were not the most skilled (the participants were not selected to be representative of any specific aspects).

As a summary of the above, one related variable confirms the validity of the computer assessment for one skill, but, for all other cases, the potentially related variables were unusable and neither confirmed nor denied validity.

Learning from this situation, the following should be done in future studies:

- Rely more on participants’ self-judgments, not only of their experience but also of their skills.
- Overcome the risk of laymen’s misunderstanding of skill definitions by pretesting their understanding.
- In addition to upraising self-judgments of larger skill entities (e.g. leadership or customer conversations), also ask about self-judgments of the small skill entities that are measured (giving feedback to employee, building up an emotional bond with an unknown customer in a first meeting and a formal situation, interviewing a customer on his or her requirements in case of a complex product). On one hand, smaller skill units might improve self-judgment, as they are more precisely defined as to what they contain (asking for requirements is a well defined activity, while customer conversation is a collection of many different skills). On the other hand, this might increase the mistakes committed by inexperienced participants when deriving self-judgments for unknown situations from known situations. Successful behaviour can be exactly the opposite in the workplace from what it is in private friendly situations. For example, while one would start positively when giving feedback to an employee, even if the criticism outweighs the praise, a trusted friend might wonder ‘Why don’t you tell me straight?’ Here, self-judgments of general leadership abilities might be more helpful, as they can be reasonably seen to reflect how well one adapts to different conversation partners.
- Assembling a set of participants comprising a more balanced and representative cohort should result in some correlations with age.

The above discussion makes it clear that it is challenging to find appropriate related variables if there are many inexperienced participants.

5.3.7 Measuring unintended constructs

Many assessment methods designed to automatically assess social skills have assessed other constructs, such as cognitive performance (see Section 3.1.1). Hence, this threat ought to be investigated carefully in the case of the PM Game. As it is a computer game featuring an artificial communication interface and using English, the threats include assessments of computer or gaming experience, cognitive performance, personality, knowledge, and English skills. Further undesired variance could stem from memory effects or undesired human influence as part of the validation process. These have already been excluded in Section 5.3.5.

The above means a statistical analysis of many variables representing attributes of the study participants or representing relationships of the study participants to the software.

Overall

A first indication can be obtained from cross-correlations among the three computer assessments:

Correlated computer assessments	Spearman's rho
Feedback versus emotional	.380** (.009)
Feedback versus requirements	.476** (.001)
Emotional versus requirements	.047 (.759)

Table 17: Cross-correlations between the computer assessments

Legend:

- Feedback: Giving feedback to employee
- Emotional: Establishing an emotional bond with the customer
- Requirements: Eliciting customer requirements

As shown by Table 17 above, there is no correlation between the computer assessments of 'Establishing an emotional bond with the customer' and 'Eliciting customer requirements'. This is an indication that the PM Game does not systematically produce extraneous variance into all its assessments. If there were a substantial variance shared by all assessments, this would also appear in the cross-correlations. Furthermore, it is particularly favourable that 'Establishing an emotional bond with the customer' and 'Eliciting customer requirements' have been assessed in the same PM Game story. Hence, this no-correlation result is an indication that the variance in the computer assessments stems from the only things that were different—the tasks and the participants' performance in completing them.

It should be noted that the above argument allows only either nothing or an indication in favour of the validity of the PM Game. An indication against cannot be derived, as it is perfectly plausible for social skills to be correlated, especially when within the same area of application (communication in business). Moreover, this is not hard evidence, as two skills could always be negatively correlated, and the negative correlation could be 'filled up' to result in zero with extraneous variance from any source. Thus, the above no-correlation result is merely presented as a weak indication that is, however, highly apparent when examining the correlations.

Threat: assessment of gaming experience and general computer usage

The PM Game skill scores could be related to the general computer usage of the participants or, more specifically, to their experience with computer games. Thus, in a questionnaire administered to the participants, the following was asked. For each question, Spearman's correlations with the computer skill assessments are provided.

Question/answer options	Feedback	Emotional	Requirements
Ich verbringe pro Woche typischerweise ... Stunden am Computer (I spend usually … hours per week at the computer) 1 = 0–1 h, 2 = 1–5 h, 3 = 5-10 h, 4 = 10–15 h, 5 = >15 h	.357* (.014)	.228 (.128)	.125 (.406)
Ich spiele pro Monat ca. ...-mal ein Computerspiel (I play a computer game… times per month) 1 = 0–1, 2 = 1–5, 3 = 6-10, 4 = 11–15, 5 = >15	.427** (.003)	.385* (.008)	-.061 (.688)
Ich habe viel Erfahrung mit Computerspielen (I have much experience with computer games) 1 = *trifft zu* (true), 2 = *trifft eher zu* (mostly true), 3 = *teils-teils* (partly/partly), 4 = *trifft eher nicht zu* (mostly false), 5 = *trifft nicht zu* (false).	-.268 (.069)	-.214 (.154)	.110 (.466)
Ich hatte eine Phase, in der ich viel Computerspiele spielte (I had a phase in which I played computer games often); same scale as above	-.276 (.060)	-.206 (.169)	.069 (.649)

Table 18: Correlations of the computer scored skills with computer and game usage

Legend:

- Feedback: Giving feedback to employee
- Emotional: Establishing an emotional bond with the customer
- Requirements: Eliciting customer requirements

It appears that gaming and general computer usage is correlated with the computer assessment. However, a closer look reveals suspiciously strong correlations for 'giving feedback to employee', mostly weak correlations for 'establishing an emotional bond with the customer', and no correlation for 'eliciting customer requirements'. This is suspicious, as all three skills use the same game interface, and the latter two even use the same PM Game story. Visualising the statistics using scatter plots reveals a classic outlier situation:

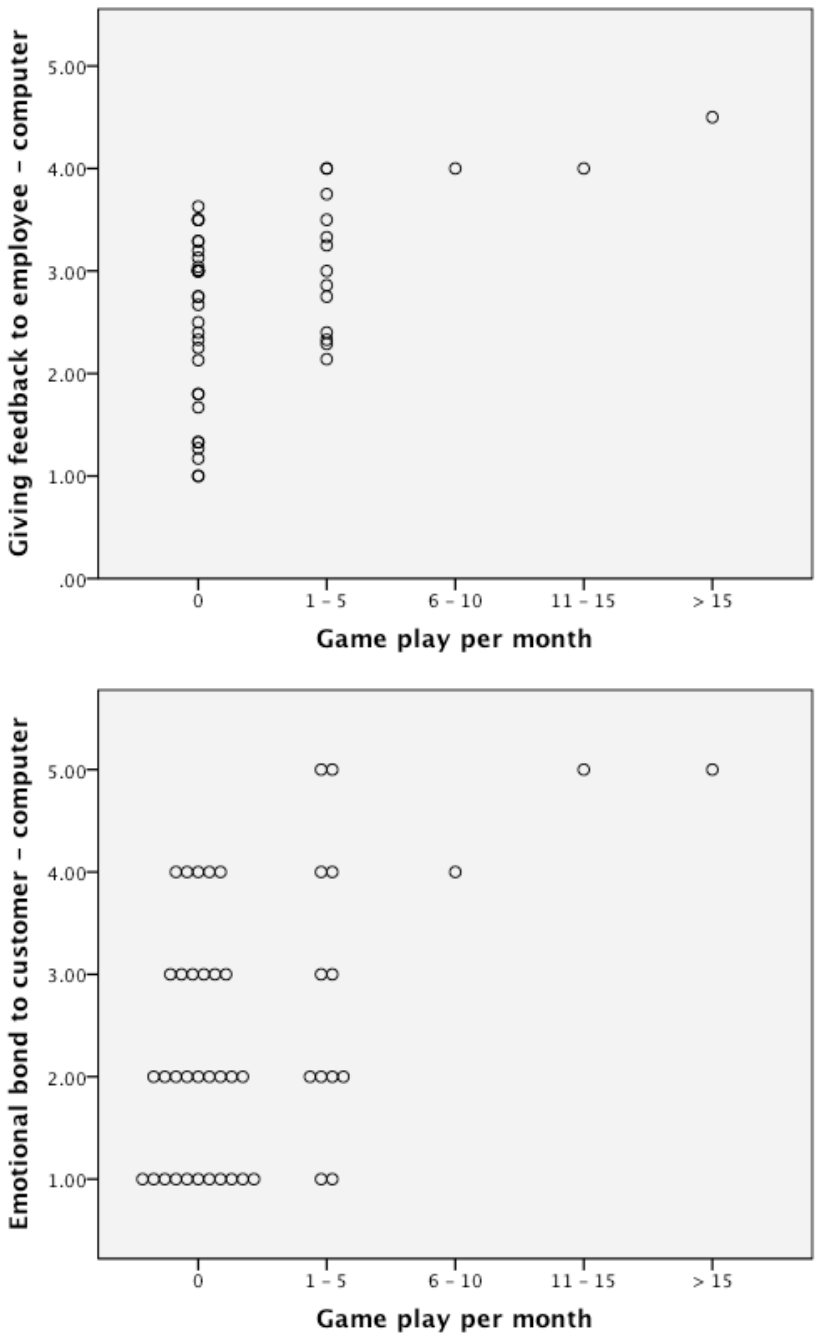


Figure 29: Scatter plots showing outliers concerning game play per month

It is to be noted that, in the second diagram, the SPSS option ‘stack identical values’ has been used because, otherwise, all the points where ‘game play per month’ is zero would fall onto each other. Using this option, the outlier situation becomes clear.

Without the three outliers, the significant correlations disappear:

Question/answer options	Feedback	Emotional	Requirements
Ich verbringe pro Woche typischerweise ... Stunden am Computer (I spend usually … hours per week at the computer) 1 = 0–1 h, 2 = 1–5 h, 3 = 5-10 h, 4 = 10–15 h, 5 = >15 h	.295 (.052)	.158 (.312)	.116 (.457)
Ich spiele pro Monat ca. ...-mal ein Computerspiel (I play a computer game… times per month) 1 = 0–1, 2 = 1–5, 3 = 6-10, 4 = 11–15, 5 = >15	.273 (.073)	.225 (.147)	-.114 (.467)
Ich habe viel Erfahrung mit Computerspielen (I have much experience with computer games) 1 = *trifft zu* (true), 2 = *trifft eher zu* (mostly true), 3 = *teils-teils* (partly/partly), 4 = *trifft eher nicht zu* (mostly false), 5 = *trifft nicht zu* (false).	-.132 (.395)	-.076 (.627)	.148 (.345)
Ich hatte eine Phase, in der ich viel Computerspiele spielte (I had a phase in which I played computer games often); same scale as above	-.179 (.244)	-.100 (.523)	.098 (.532)

Table 19: Correlations of computer scored skills with computer and game usage, outliers removed

Without the two further ‘best supporters’, there is no more correlation:

Question/answer options	Feedback	Emotional	Requirements
Ich verbringe pro Woche typischerweise ... Stunden am Computer (I spend usually … hours per week at the computer) 1 = 0–1 h, 2 = 1–5 h, 3 = 5-10 h, 4 = 10–15 h, 5 = >15 h	.243 (.121)	.114 (.477)	.085 (.598)
Ich spiele pro Monat ca. ...-mal ein Computerspiel (I play a computer game… times per month) 1 = 0–1, 2 = 1–5, 3 = 6-10, 4 = 11–15, 5 = >15	.172 (.275)	.145 (.367)	-.196 (.220)
Ich habe viel Erfahrung mit Computerspielen (I have much experience with computer games) 1 = trifft zu (true), 2 = trifft eher zu (mostly true), 3 = teils-teils (partly/partly), 4 = trifft eher nicht zu (mostly false), 5 = trifft nicht zu (false).	-.127 (.425)	-.101 (.531)	.126 (.431)
Ich hatte eine Phase, in der ich viel Computerspiele spielte (I had a phase in which I played computer games often); same scale as above	-.185 (.242)	-.129 (.422)	.074 (.643)

Table 20: Correlations of computer scored skills with computer and game usage, 5 points removed

As an intermediate summary, the correlations seen in Table 18 are caused by a small number of data points, but are these data point errors produced by either the method or the PM Game or do the study participants simply have excellent skills and high game and computer affinities? This question can be answered by examining the relationship between the game and computer affinities and the skill scores produced by the human assessors. The scatter plot reveals the same three outliers:

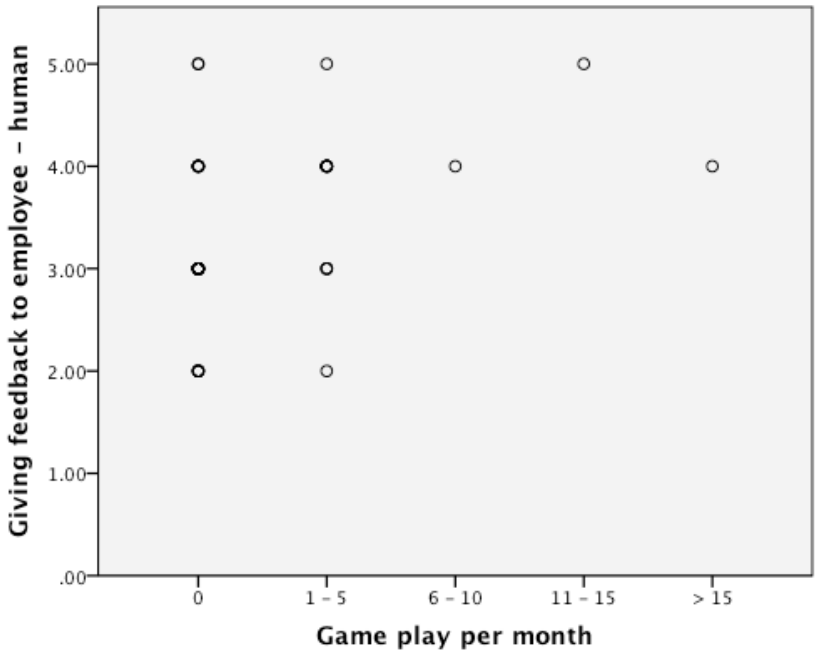


Figure 30: Human assessments have the same three outliers

The correlations are also similar:

Question / answer options	Feedback	Emotional	Requirements
Ich verbringe pro Woche typischerweise ... Stunden am Computer (I spend usually … hours per week at the computer) 1 = 0–1 h, 2 = 1–5 h, 3 = 5-10 h, 4 = 10–15 h, 5 = >15 h	.227 (.125)	.231 (.118)	.041 (.785)
Ich spiele pro Monat ca. ...-mal ein Computerspiel (I play a computer game… times per month) 1 = 0–1, 2 = 1–5, 3 = 6-10, 4 = 11–15, 5 = >15	.389** (.007)	.186 (.210)	-.082 (.582)
Ich habe viel Erfahrung mit Computerspielen (I have much experience with computer games) 1 = *trifft zu* (true), 2 = *trifft eher zu* (mostly true), 3 = *teils-teils* (partly/partly), 4 = *trifft eher nicht zu* (mostly false), 5 = *trifft nicht zu* (false).	-.226 (.126)	-.100 (.506)	.139 (.353)
Ich hatte eine Phase, in der ich viel Computerspiele spielte (I had a phase in which I played computer games often); same scale as above	-.312* (.033)	-.268 (.068)	-.004 (.979)

Table 21: Correlations of human scored skills with computer and game usage

Removing the same five study participants as above results, again, in no correlation:

Question/answer options	Feedback	Emotional	Requirements
Ich verbringe pro Woche typischerweise ... Stunden am Computer (I spend usually … hours per week at the computer) 1 = 0–1 h, 2 = 1–5 h, 3 = 5-10 h, 4 = 10–15 h, 5 = >15 h	.124 (.432)	.213 (.177)	.106 (.503)
Ich spiele pro Monat ca. ...-mal ein Computerspiel (I play a computer game… times per month) 1 = 0–1, 2 = 1–5, 3 = 6-10, 4 = 11–15, 5 = >15	.222 (.158)	.128 (.420)	.021 (.895)
Ich habe viel Erfahrung mit Computerspielen (I have much experience with computer games) 1 = *trifft zu* (true), 2 = *trifft eher zu* (mostly true), 3 = *teils-teils* (partly/partly), 4 = *trifft eher nicht zu* (mostly false), 5 = *trifft nicht zu* (false).	-.148 (.350)	-.013 (.935)	.005 (.976)
Ich hatte eine Phase, in der ich viel Computerspiele spielte (I had a phase in which I played computer games often); same scale as above	-.278 (.075)	-.215 (.172)	-.146 (.357)

Table 22: Correlations of human scored skills with computer and game usage, 5 points removed

The above evidence shows that a few participants had, simultaneously, high skills and high computer and game affinity. This caused the impression of bias—that the computer assessment might prefer people with computer or game experience. Thus, there is no indication that the PM Game scores were contaminated by variances from computer or game affinity.

While Spearman's rho is a robust measure of dependence, with an increasing number of outliers, the so-called 'breakdown point' can be reached. The presence of several study participants with both high skills and high computer or game affinity is readily explained by the fact that participation in the study was voluntary, and people with high game affinity and leadership skills could have been attracted to it. Further studies should be carefully advertised, or studies should not use volunteers in order to avoid self-selection effects.

Threat: assessment of cognitive performance (cognitive load)

In order to test how much influence cognitive performance had on the computer scores, participants were asked questions related to their cognitive load when playing the PM Game. The capacity to perform high cognitive load tasks is one aspect of cognitive performance and is related to intelligence, as both use working memory capacity as a central element (Pass et al., 2003; Convay et al., 2003). However, many studies use an intelligence test to determine whether a variable contains variance from cognitive performance. Hence, before reviewing the correlations, the question of whether measuring cognitive load was the right approach will be examined. The PM Game presents the user with the following:

- A new virtual world, with many elements and procedures that must be learnt (e.g. where to click to communicate). It was a precondition that participants did not know the game beforehand.

- The virtual world and its social situations are designed not to challenge logical, numerical, spatial, or verbal intelligence: it is a simple 2-dimensional screen, with simple English sentences, using no comparisons or tricky differences in meaning. The stories do not contain any logical difficulties or predefined alternatives that had to be compared and analysed for advantages and disadvantages, such as, for example, in multiple-choice tests.

Thus, the situation was dominated by challenges from many new things at once and not by challenging typical intelligence test areas, such as logical, numerical, verbal, or spatial intelligence.

Last but not least, intelligence tests are time consuming: a full intelligence test, such as Mensa's membership test, lasts 1.5 hours, but even short intelligence tests for logical, verbal, numerical, or spatial intelligence require 10 minutes or more for each part (e.g. IST-2000R, Liepmann et al., 2007). This additional time has not been required from the participants in this study because there was a risk of participant demotivation. Participant demotivation during extensive testing may produce serious drawbacks in data quality. Indeed, this probably happened to a small extent in this study: one participant said only 'hello' and left in one of the stories and thus did not really play the game. While a normal assessment centre study may request more effort, it also provides more value to the participant. Even full days of assessment centres can be motivated when presented to the participants as assessment centre training (e.g. Kleinmann, 1997, pp. 197-215). Such options to motivate participants are, however, not available in the PM Game, as it is not a preparation for current job applicant testing methods. Moreover, the PM Game is in particular danger of incurring participant demotivation, as it is played alone, with no assessors or other people watching.

To measure cognitive load, appropriate self-report questions were developed. The subject's self-report is the established means of surveying cognitive load; compare the NASA Task Load Index (Hart & Staveland, 1988; Hart, 2006). To best uncover the cognitive load caused by the PM Game, specific questions were developed matching the current test situation in which participants first face the PM Game. In addition, a general question about the ability to handle new situations was asked, as well as several questions that survey specific sources in the PM Game from which cognitive loads could arise.

Question/answer options	Feedback	Emotional	Requirements
Questions regarding the game experience:			
Ich habe keinen Überblick über das viele Neue (I have no overview of the many new things) 1 = *trifft zu* (true), 2 = *trifft eher zu* (mostly true), 3 = *teils-teils* (partly/partly), 4 = *trifft eher nicht zu* (mostly false), 5 = *trifft nicht zu* (false)	.085 (.570)	.161 (.285)	-.174 (.248)
Das Spiel ist zu viel auf einmal, ich bräuchte länger um mich daran zu gewöhnen (The game requires too much at once; I would need more time to get used to it); same answer scale as above	.084 (.575)	.298* (.044)	-.111 (.462)
General question of ability to deal with c.l.:			
Ich kann mich in komplexe neue Dinge schnell einarbeiten (I can learn complex new things fast)	-.172 (.248)	-.081 (.592)	-.127 (.400)
Specific sources of possible cognitive load:			
Die Spieloberfläche ist … (The game interface is …)			
frustrierend–überzeugend (frustrating–convincing); analogous answer scale, with 9 choices between the poles.	-.034 (.820)	.106 (.485)	-.261 (.080)
schwierig–einfach (difficult–easy); same answer scale as above	.025 (.867)	.130 (.387)	-.033 (.828)
starr/steif–flexible (inflexible–flexible); same answer scale as above	.059 (.694)	.010 (.946)	-.183 (.224)
Den richtigen Sprechakt zu finden… (To find the right speech act …); same scale			
war schwierig–war einfach (was difficult–easy); same answer scale as above	.150 (.315)	.227 (.130)	.006 (.967)
ging langsam–ging schnell (went fast–went slow); same answer scale	-.091 (.544)	.351* (.017)	-.162 (.281)
Die Spiel-Geschichte zu verstehen war … (To understand the game story was …)			
schwierig–einfach (difficult–easy); same answer scale as above	.162 (.278)	.135 (.371)	.002 (.992)

Table 23: Correlations of computer scored skills with questions relevant to cognitive load

Table 23 shows that there is little correlation with cognitive load. Some influence in the case of 'establishing emotional bond with customer' is likely, but the other skills show no correlations with cognitive load.

It should be noted that NASA-TLX was not used as a general and established means of surveying cognitive load because most NASA-TLX items survey sources of cognitive load irrelevant to the present situation (e.g. physical effort, time pressure, stress level, fatigue). The remaining few questions are applicable to the present situation but only generally (e.g. 'Was the task demanding or frustrating?' and 'Does the participant think he performed well?').

Generally, any standardized test will have an advantage when the comparability of one task's cognitive load with that of another is required. This is, however, not the purpose here. The purpose is to discover the cognitive load arising from a specific assessment tool, the PM Game. For this, it is more appropriate to ask specific questions covering the different cognitive load sources that could arise.

Threat: influence of English skills on the test scores

Study participants were requested to have English skills to 'understand written dialogues quickly'. The PM Game was pretested with similar users to ensure that English problems did not occur. However, as English was not the mother tongue of the participants, English skills may have influenced the test results. In other words, it is possible that the computer scores measured English skills to a degree. To determine this, the participants were asked the following question:

Question/answer options	Feedback	Emotional	Requirements
Ich vermute, dass ich wegen dem Englisch etwas falsch gemacht habe (I think that I have committed mistakes because of the English) 1 = *trifft zu* (true), 2 = *trifft eher zu* (mostly true), 3 = *teils-teils* (partly/partly), 4 = *trifft eher nicht zu* (mostly false), 5 = *trifft nicht zu* (false).	.247 (.094)	.233 (.120)	.147 (.330)

Table 24: Correlation of participants' English language issues with computer skill scores

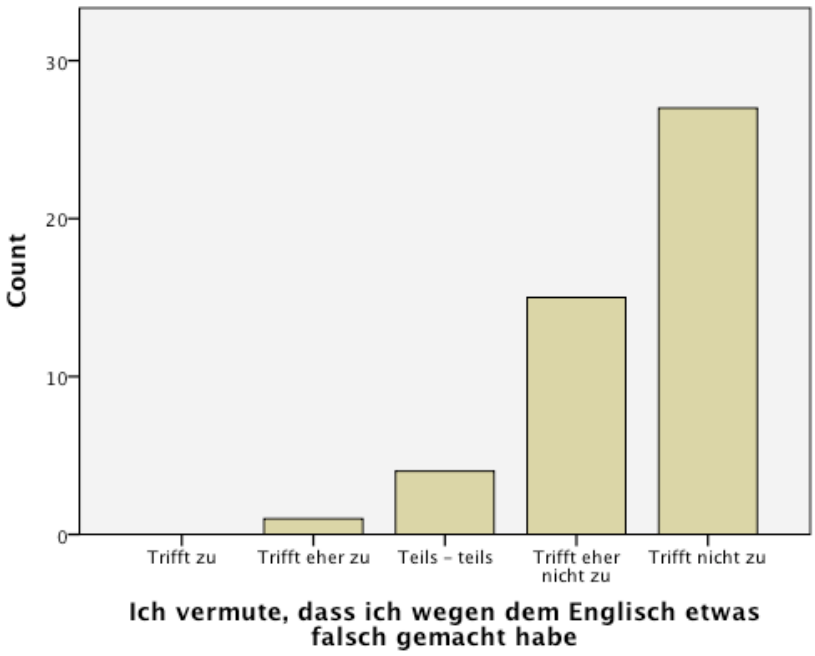


Figure 31: A small number of participants felt hindered by their English skills (translations above)

As shown, a small number of participants felt that their English was a hindrance. There is no significant indication that this explains variances in the test results. However, Table 24 appears to indicate some non-significant correlation. Thus, it might be that, for the larger participant numbers, some variances in the test scores could be significantly explained by differences in English skill.

Some contamination with English skill variances is not an issue for the PM Game as a method of social skills assessment, because English was used merely as a convenience to save development effort. In actual industry usage, the game would be provided in the participant's native language. For further scientific enquiry, it will be important to know that administering the game in English to non-natives is not a hindrance. However, in subsequent studies, participants must be required to have good English skills.

Threat: assessment of personality or of knowledge

Two basic questions have been administered regarding personality—whether participants like to work on abstract tasks or with people in conversation.

Question/answer options	Feedback	Emotional	Requirements
Ich arbeite gerne an abstrakten Aufgaben (I like to work on abstract tasks) 1 = *trifft zu* (true), 2 = *trifft eher zu* (mostly true), 3 = *teils-teils* (partly/partly), 4 = *trifft eher nicht zu* (mostly false), 5 = *trifft nicht zu* (false)	-.285 (.052)	-.136 (.368)	-.186 (.216)
Ich arbeite gerne mit Menschen im Gespräch (I like to work with people in conversation); same scale.	.074 (.619)	.011 (.941)	-.144 (.340)

Table 25: Spearman's rho of personality questions vs. computer assessments

The above data indicate no, small, and non-significant correlation of skill scores with the two questions. The author does not claim that these two questions could replace a full personality test. However, these two questions have been chosen as typical and business-relevant personality questions:

- The main dimensions of the Grid leadership theory are task orientation and people orientation (Blake & Mouton, 1964 & 1985, McKee & Carlson, 1999).
- Questions about (for example) whether a person prefers reading books or socializing with people are typical elements of personality tests to determine scores on the extraversion-introversion scale of many popular personality tests, such as the Myers-Briggs type inventory (1980 & 1995) and the 'NEO-FFI' (Costa & McCrae, 1992).

A full personality test has not been administered in order to avoid overtasking the study participants.

Regarding knowledge, it is highly unlikely that participants did not possess active or passive knowledge of the social rules that were tested. Many were simple, well-known social rules that most study participants have probably encountered several times and with which everyone would generally agree, such as starting feedback positively (at least in formal situations), being particularly attentive when talking to a customer, providing a conversation partner with utterances such as 'hm', 'yes', 'I see' to show that one is paying attention, and asking questions in order to find out what a customer wants. However, as can be judged from the actual course of role-plays and PM Game stories (see the accompanying CD, directory: 'Scoring of assessment centre performance'), even these obvious and well-known rules of approaching customers and giving feedback were incorrectly done by many participants, both in the role-plays and in the PM Game. This is exactly the difference between knowledge and behaviour.

5.3.8 Discriminant validity

An assessment method for skills has discriminant validity if correlations with the skills it claims to assess are larger than correlations with other skills. According to Table 25, this considtion is fulfilled for the computer assessments of 'giving feedback to employee' and 'establishing emotional bond with customer'. In case of 'eliciting customer requirements' this is not the case. If the computer based assessment method as a whole has discriminant validity, cannot be concluded from assessing three skills. For this more research is needed. However, this is likely to be the case because computers can be assumed to be not subject to the halo effect (Section 3.2.2). Also, the computer assessments for 'eliciting customer requirements' are based on substantially less behavioural markers compared to the other two skills. Hence, there is potential that increasing the behavioural markers discriminant validity could be achieved.

	Feedback computer	Feedback human	Emotional computer	Emotional human	Requirem. computer	Requirem. human
Feedback computer	1.000	.701***	.380**	.214	.476***	.046
	.	.000	.009	.149	.001	.758
	47	47	46	47	46	47
Feedback human	.701***	1.000	.320*	.387**	.447**	.131
	.000	.	.030	.007	.002	.381
	47	47	46	47	46	47
Emotional computer	.380**	.320*	1.000	.511***	.047	-.090
	.009	.030	.	.000	.759	.552
	46	46	46	46	46	46
Emotional human	.214	.387**	.511***	1.000	.214	.266
	.149	.007	.000	.	.153	.071
	47	47	46	47	46	47
Requirements computer	.476***	.447**	.047	.214	1.000	.383**
	.001	.002	.759	.153	.	.009
	46	46	46	46	46	46
Requirements human	.046	.131	-.090	.266	.383**	1.000
	.758	.381	.552	.071	.009	.
	47	47	46	47	46	47

Table 26: Cross-correlations of human and computer assessments

Legend:

- Feedback: Giving feedback to employee
- Emotional: Establishing emotional bond with customer
- Requirements: Eliciting customer requirements
- The entries in the table fields are: correlation coefficient, significance, N.

5.4 Validation study: interpretation of the main result

How good are the results presented in Section 5.3.1? How large could the correlations maximally be? If the social skills assessment of the PM Game is as good as that of the human assessors, then the correlations of the PM Game and the human scored role-plays will be maximally as good as the correla-

tions between carefully constructed parallel versions of the human-scored assessment methods. This benchmark, called the 'retest reliability of parallel versions', is not readily available, however.

In what is probably the most cited retest study, Moses (1973) measured a correlation of r = .73 for the scores of subjects taking part in two similar assessment centres shortly after each other. This correlation is sometimes regarded as an estimate of the retest reliability of assessment centres (compare e.g. Kelbetz & Schuler, 2002). The result of Moses' study is unfortunately not comparable to the case in this thesis for the following reasons:

- The reported correlation is that of the overall scores the participants achieved in the two assessment centres (while, in this thesis, the scores for individual assessment centre tasks are compared)
- There were no role-play tasks in the two assessment centres. They consisted of written tests, in-baskets, and leaderless discussions (while this thesis uses role-playing)
- The article also reports correlations between the assessment centre tasks (many different values spanning the range from -.34 to .73), but there is no statement on how similar the tasks are, and it is not stated that the tasks would have been designed to assess the same.

A similar result is reported by Kleinmann (1997, pp. 197-215), who found two short assessment centres executed on the same day to be correlated by r = .65. The first short assessment centre consisted of a group discussion without assigned roles and a presentation. The second consisted of a leaderless group discussion with assigned roles and a role-play. Thus, the assessment centre exercises were different and not parallel versions of each other. Furthermore, no correlation of task scores was reported. However, the assessment centres as a whole had been constructed to assess the same. This result of Kleinmann's is more comparable with the results of this thesis' validity study, as the overall correlated scores were the aggregation of only two assessment centre tasks (one in this thesis), and the assessment centre tasks were all behavioural (as in this thesis).

Funke & Schuler (1998) reported a correlation of r = .65 for the overall score of six role-play tasks, once administered to the participants using video stimulus and once using oral stimulus. The oral stimulus was a description of a work situation (i.e. context, persons, actions) in a few sentences analogous to a situational question. One week elapsed between tests. Similar to the setting of this thesis, the task content was 50% customer communication and 50% communication between a subordinate and a superior. The main difference with the results in this thesis is that only the correlation of the overall score of the six role-play tasks was reported. No information is available on the scores of the individual tasks, which would have been more comparable.

Finally, Kelbetz & Schuler (2002) report r = .41 for the retest validity of their assessment centre after an average time-span of two years between two tests with the same assessment centre.

It can be concluded, then, that the best comparable retest study is in Kleinmann (1997), followed by Funke & Schuler (1998), suggesting that it may be reasonable to assume that the correlation of .701 in this study (computer assessment versus human assessment of 'giving feedback to employee') is as good as retest correlations can maximally be and that, even if an assessment by trained human assessors were used for both assessments in carefully constructed research-grade assessment centres, the retest correlation would probably not be better than this. The result of .511 (computer assessment versus human assessment of 'establishing emotional bond with customer') approaches this. A further hallmark of social skills assessment by trained human assessors is that the scores contain no or little extraneous variance, stemming, for example, from cognitive performance. Accordingly, the correla-

tion results presented in Section 5.3.7 indicate that the computer scores contain no or little extraneous variance.

In conclusion, the main results of the validation study can be summarized as follows:

In this validation study, the computer assessment of the verbal aspects of one social skill has been as good as the human assessment. For a second social skill, the computer assessment came close to that result.

As a remark, there is some indication to assume that this quality of assessment can be achieved systematically: comparing the results of this benachmark comparison with the number of behavioural marks used for the computer assessment of the three skills (Table 7 in Section 5.3.1), the quality of assessment strongly increases with the number of behavioural markers used. Clearly, further research is needed to substantiate this hypothesis.

The above statements must be made with the following caveats:

- In this thesis' validation study, only the verbal aspects were assessed both in the human scored role-plays and in the computer game test. Nonverbal aspects are highly relevant for these skills and for social skills in general.
- The data used for the benchmarking in this section only approximate the case of this study and do not fit perfectly. Better comparison data are not available. Research on parallel assessment centre tasks is needed. This was confirmed both by the author's literature analysis as well as by Brummel et al. (2009).

5.5 Examination of feasibility using independent research

In this section, the feasibility of the proposed method will be examined using the research results of other authors. Verification of feasibility based on independent data is always advisable when claiming a solution for a long-standing problem. In this particular case, doubt may arise concerning the following issues:

- It is not apparent whether social skills can be assessed based on written human–computer dialogues if all non-verbal expression by the user is ignored. Nonverbal cues help communicators interpret the subtle meaning behind others' messages (Archer & Akert, 1977). Although researchers do not agree on what proportion of the meaning of a message is communicated nonverbally, it is generally accepted that in many conversations more information is transmitted nonverbally than verbally (e.g. Mehrabian & Wiener, 1967).
- To process communication, the PM Game (as well as many other communication games) uses internal representations that are even more limited than written free text.

Hence, in this section the following will be examined:

1. Is it possible to assess social skills using written dialogues (zero nonverbal information)?
2. Are communication representations in communication games suitable for social-skills assessment?

These questions will not be answered focussing solely on the PM Game, but generally, focussing on written dialogues in various communication games.

5.5.1 Can social skills be assessed based on human-computer text dialogues?

To study this issue, research by Funke & Schuler (1998) will be used. Research on the assessment of social skills using written dialogues is rare. Funke & Schuler's was the only study on this subject discovered by the author after reviewing the literature in detail. In this section, first, a brief summary of the study will be presented. Then the conclusions will be discussed.

Funke & Schuler (1998) investigated the relationship of validity to the stimulus format and response format of tests for assessing social skills. They compared the suitability in assessing social skills of different response formats: situational judgment tests (i.e. multiple choice questions), free-text answers to preformulated test questions, and oral responses (situational interviews and role-plays). Moreover, they compared different stimulus formats: all tests came in oral and video variants. All these test variants contained the same content: six situations requiring occupational social competency. These situations have been used by a German regional bank as part of an assessment centre.

- Customer courtesy in a routine situation
- Customer wishes conflict with regulations
- Customer with risky financing needs
- Subordinate must convince superior to delegate a consulting task to the subordinate
- Subordinate must inform superior about an error
- Superior must justify higher-order organizational goal requirements to subordinates

For the 75 participants, assessment by situational judgment test (SJT) did not significantly correlate with the social skills shown in the role-play assessments ($r = .13$ in the case of oral stimulus and $r = .17$ in the case of video stimulus). However, assessment by free-text answer did correlate ($r = .36^{**}$, $p < .01$ and $r = .37^{**}$, $p < .01$). Thus, the stimulus format did not play any substantial role for the validity of the assessment. Concerning the response formats, written free-text answers resulted in significant validity, while multiple choice did not.

The results of Funke & Schuler (1998) show that it is possible to assess several social skills using written answers. A difference between their study and the PM Game is that Funke & Schuler only asked study participants to produce one single written text per situation (in a paper-and-pencil test) and not to interact (with the computer) using repeated answers. In this regard the PM Game – and games in general – have an advantage: repeated interaction is possible. Thus, these interactions are more similar to real conversations. (Remark: this advantage might explain the better correlations in case of the PM Game because the difference lies in the response format. The results of Funke & Schuler, as well as other research on stimulus and response formats in assessment methods (Karkoschka, 1998), shows that the fidelity of the response format is what matters for the assessment of social skills.)

5.5.2 Are communication representations in games suitable for social skills assessment?

The PM Game, and current communication games in general, process communications by performing transformations to internal representations. Thus, this section will address the question of whether it is possible for these internal representations to contain information sufficient for the assessment of social skills.

Communication representations in games

Although there is no standard way how communication is processed in games, the major steps in a full-fledged state-of-the-art system are usually those depicted in Figure 32.

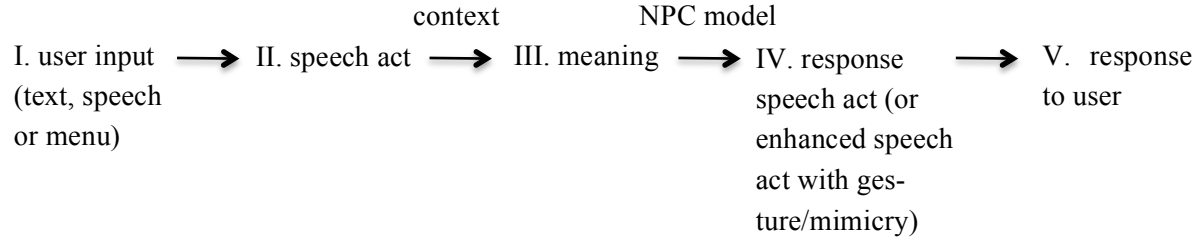


Figure 32: Typical major phases of communication processing in games

Legend: 'context' denotes the current game state, including the communication history. 'NPC' denotes the nonplayer character that produces the reaction.

The above diagram has been developed by comparing various communication games. The full process is implemented, for example, in Façade (Mateas & Stern, 2003) and in IDTension (Szilas, 2007). In both cases, the user input (Phase I) is mapped onto a fixed set of speech acts (Phase II), which is then itself mapped (depending on the context, i.e. the current state of the game, including previous communications) onto a fixed set of meanings (Phase III). Though being fixed, speech acts and meanings can be parameterised to result in a large number of possibilities (e.g. the verb or object in a speech act can be flexible). Steps IV and V are presented in order to give a complete picture, but do not play a role in the argument.

Figure 32 should not be understood as a uniform method of processing in games but rather as a concept in whose terms many processing variants can be explained. This is for example the case in the PM Game: here, Phases I and II are identical, as the user directly produces a speech act via his or her menu choice. Phase III is not explicitly represented: from Phase II, using infoobjects and speech-act rules described in Section 4.8, Phase IV is computed and subsequently displayed to the user as text in speech bubbles (Phase V). The assessment scores are computed using the dialogues (Phase I and Phase V). As part of this process, behavioural markers are identified. These define the meaning of an action in a social situation (Phase III); see Section 4.6. However, as described for IDTension and Façade above, the set of meanings is fixed. (A fixed set of behavioural markers is used. They have been crafted in advance). This fact will be important for the following discussion.

A fixed number of predefined meanings suffice for social-skills assessment

The examination of the processing phases of communication in games raises the following question: is assessment of social skills possible if the written text of human–computer dialogues is mapped to a fixed set of meanings (behavioural markers)? Based on common ideas about the complexity of human behaviour and social skills, it seems reasonable to assume that this is not possible. The behavioural checklist method (e.g. Henessy et al., 1998) used in state-of-the-art assessment centres, however, uses a fixed set of predefined behavioural markers, although the input there is not written text but spoken interaction. Regarding written input, the research of Funke and Schuler will again be relevant, as it

uses a method similar to computer processing. Their procedure for assessing free-text answers is described in Schuler (1993): using previously developed checklists, human assessors verify the presence or absence of certain predefined meanings in the free-text answers. The amount of the latter is the score. Thus, Funke & Schuler (1998) have provided evidence that information on the presence or absence of a fixed choice of predefined meanings in a free-text answer to a well-designed question is sufficient to assess social skills. This is equivalent to the communication representation III in Figure 32. Thus, this representation is capable of containing the information needed for the assessment of social skills.

5.5.3 Summary and conclusion

The PM Game and the other communication games examined above map written text-only communications to a fixed set of meanings (behavioural markers). Funke & Schuler (1998) provide empirical evidence that such a mapping (carried out by humans) was able to assess social skills. This confirms the feasibility of core elements of the method proposed in this thesis for social-skills assessment.

This feasibility analysis, however, considered neither computer-based mapping processes between data formats nor the human–computer communication interface. In Section 3.1.1, evidence has been presented that using multiple choice as the input method it is probably not possible to assess social skills. Free-text input has been confirmed by Funke & Schuler to be capable of containing the required assessment information; however, it is difficult for computers to carry out the required mapping processes. In conclusion, neither multiple choice nor free text appears to be a feasible method of user communication input for computer-based social-skills assessment using currently available technologies. Menu-based communication remains a possible option for computer-based social-skills assessment.

5.6 Analysis of extensibility

With regard to the use cases presented in Chapter 2, in order to be suitable for industrial application, the PM Game must be capable of assessing not only three small skills but a multitude of them. Furthermore, if a particular skill can only be assessed by a particular story, then test-takers will get used to these stories, resulting in artificially high scores. For these reasons, extensibility is a central criterion for the PM Game and for the computer-based assessment method it implements.

The best proof of extensibility is to extend the game. In the case of the PM Game, this means not only to develop but also to successfully validate the assessment stories. However, until this work is carried out, the following analysis should provide insight into where the opportunities and difficulties lie.

5.6.1 Extension tests and authoring guidelines

The first extensibility test of the PM Game was carried out by Ito (2009). It concluded that students in various fields were able to create game stories. However, this study did not focus on developing game stories for the purpose of assessment. The goal was to create any playable stories. To gain insight into the creation of game stories for assessment, the author of this thesis supervised three people creating PM Game stories: a secondary school student, a psychology student and a software developer. These experiences were documented in the form of the guidelines in Appendix 7.7 in order to make the creation of assessment stories a learnable and repeatable process. Each recommendation given there corresponds to a mistake that has actually been committed or a positive experience that proved useful.

Creating stories for the purpose of assessment is a skill on its own that needs to be learnt. The major pitfalls to avoid in this process can be summarised as follows:

- Stories should not to be invented by free imagination and then made to fit into the game, but instead need to be crafted around the pool of available sentence stubs.
- Presenting explicit choices needs to be avoided. While the game is designed to create tests in which the options are not explicitly presented (Section 4.2), the game authors found a means to circumvent this and return to common multiple-choice tests: they embodied the choices as items. Items are intended to present the elements of the social situation (Section 4.3).
- The difference between stories in media such as films, theatre and books, on the one hand, and computer games on the other needs to be learnt: while in the former, the characters always do what the author imagined, in the latter protagonists (users) can say anything, anytime permitted by the interface.
- Developing quality stories is highly time consuming when one has no experience with the situation modelled by the story. This should be avoided if possible.
- A substantial amount of time needs to be sacrificed to user testing.
- Development is an iterative process. This may be surprising, as assessment stories are small (they typically involve around 20 user actions in total).

The processes of creation of the three stories used for the validation study (Section 5.2) and supervising the authors also led to positive lessons:

- The small pool of currently 38 sentence stubs seems to be sufficient for a variety of stories. Though a new story might require a new sentence stub, the demand for several new sentence stubs appears to be avoidable by careful adjustment of story content while still addressing the skill that should be tested. For example, a sentence stub such as '[…] is […]' can be used to form a large variety of statements. Thus, the need of users to express statements will be covered to large extent by this single sentence stub.
- The skill of authoring stories appears to be learnable.
- Assuming that the author has experience in the subject field of his story, the amount of time needed to author one assessment story and learn the authoring process is currently estimated to be 30 full working days, not counting the validation process (Appendix 7.7).

As a disclaimer, the insights reported above are only preliminary conclusions from work in progress. They should, however, represent the best insight currently available, and they certainly will evolve to some extent as further research is carried out.

5.6.2 Bottlenecks of extension

The purpose of this section is to present issues that may arise when increasing the number of stories and/or of assessed skills in the PM Game. The focus will be on the difficult issues where the path of the solution may not be clear. Using the practical extension test reported in the previous section, the analysis of use cases in Chapter 2, the description of the design in Chapter 4, and the previous research discussed in Section 5.1.3, the following issues have been found:

- More stories will require more sentence stubs in the communication menu. The PM Game, however, relies on the assumption that only a limited number of sentence stubs is needed.
- There are skills that are probably or certainly not assessable using this approach.

Examples and reasons for these issues, their potential impact, and potential ways to cope with them will be analysed in the following sections.

As a remark, many more issues exist in this process, some of which may require substantial work. However, it is methodically clear how to resolve them. For example, an increasing library of visual icons will be needed to model the many social situations represented; also, means to navigate the stories are needed, and so on. The most challenging of these questions are how to aggregate skills into larger skill entities and how to utilize assessment information that has been generated in different stories for the same skills.

5.6.3 Issue 1: More stories require more sentence stubs

There is one piece of content that needs to be shared across stories: the sentence stubs that are available for the user to communicate with. Sentence stubs need to be the same across stories, otherwise the users will browse through the stubs to find out what is possible in the current story and, even worse, try to infer from the sentence stubs what the expected solutions may be (cf. Section 4.2). From the experiences given in Section 5.6.1, the currently most probable expectation is that the pool of sentence stubs will grow moderately with the number of stories. Fortunately, this is the only such resource that needs to be shared across stories. However, this single issue has a number of consequences. These will be presented in this section, along with suggestions on how to resolve them.

Resolvable issues

If there are too many sentence stubs, they may overwhelm novice users. This can be resolved simply by offering first only a moderate number of sentences and then adding further ones after a user has played several stories.

A more serious issue is that adding more sentence stubs implies that all previous stories need to be adapted and validated again: there needs to be an answer to all sentences the user may form, including the new ones. The always available ‘speech-act answers’ (general, not situation-specific answers; see Section 4.8.2) may or may not fit the situation. If they do not fit, one or several info-object answers need to be crafted. In both cases, new paths of user action may result. These paths may influence the assessment of the behavioural markers in the story, and the story will need to be validated again. This causes substantial effort (see Appendix 7.7). Thus, increasing the pool of available sentences should happen only rarely, collecting well-considered extensions into one update. If a larger package of stories has been collected and validated, there need to be a package-specific version of the sentence stubs. As long as users are unlikely to be confronted with different packages at the same time or within a short time, this should be acceptable.

The risk of reaching a maximum feasible amount of sentence stubs

Is there a maximum feasible number of sentence stubs? This issue constitutes an even more severe threat than the validation problem mentioned before (which can be resolved by packaging). Hence, in this section, the problem and the current situation will be analysed in detail in order to draw conclusions regarding ways to extend the game.

Having many sentence stubs does not represent a usability problem per se, because they are hidden behind the main menu: in the initial state of the menu, sentence stubs are not visible in order to prevent the users from frustrated browsing.

Figure 33: The initial state of the communication interface

The limiting factor is the number of sentence stubs that become visible during the user process of forming communications. The latter consists of a click on any element of the main menu plus a click on any object, that is, any icon visible on the screen, in any order (Section 4.5). The respective numbers of sentence stubs visible after these clicks is displayed in Table 27.

	Start/end	React	Get info	Give info	Let do	I do
Visible after click on menu	8	6	6	14	1	3
Visible after click on object	7					
Visible after clicks on menu and object	8	6	4	2	1	3
Accessible by browsing only	8	5	1	10	0	3

Table 27: Number of sentence stubs in various situations

The communication interface can display seven sentence stubs without scrolling. According to the previous experience with users of the PM Game, this is an easily manageable amount. Reviewing Table 27, there is a problem with the main menu entry ‘give info’ (14 sentence stubs). Unfortunately, there are sentence stubs to which it is not possible to navigate using the second click on an object in the social situation displayed. For example, ‘May I provide a summary?’ is a sentence under the main menu entry ‘give info’, and has no gap where an object from the social situation could be filled in. Thus, this sentence needs to be accessed by browsing. There are 10 such sentences under the main menu entry ‘give info’. That the game works despite this usability issue should not be taken as a guarantee that it is without risk. In fact, reviewing log files of game-stories played through, miscommunications can be observed in the cases of several users. By observing the users directly, it was seen that scrolling menus up and down several times did occur. Thus, there is a need not to increase the burden that this usability problem represents.

While there is a risk that user communications will be altered by the usability defect of hard-to-find sentence stubs, several additional studies would be needed to quantify this risk exactly. The information needed would be an estimate of the breakdown point: the point when this usability issue starts to affect assessment validity. As long as such studies are not available, it must be assumed that this is a considerable risk, as it affects the most sensitive aspect of the game: the assessment works only as long as the means of the game allows sufficiently realistic communication.

Possible methods to address the issue:

- Packaging as described above will also limit or even resolve this issue. The question is if a reasonable package for an application field can be constructed without overwhelming the user with too many sentence stubs. For the purposes of job assessment (Use Case 1 in Chapter 2), this can reasonably be assumed to be feasible: an assessment centre typically contains about 5–10 tasks. Thus, 5–10 stories in one package would suffice for this purpose. A job assessment does not have to cover all relevant abilities; a selection is sufficient. For the purposes of detailed feedback, for example on leadership abilities, in order to identify weaknesses and induce learning, it would be desirable to have more stories to compute a skill profile. It must be determined by future research whether this is feasible using packaging alone.
- An additional menu could be introduced. Currently, in order to produce one sentence, the user typically needs to make four clicks (main menu, icon of a situation element, sentence, checkmark button). Adding a further submenu would add one click. This appears to be a feasible option as long as this submenu is stable, so that the user can get used to it. Such a submenu needs to be found and tested, with careful usability evaluation. An example (merely to explain the idea) could be 'past, present, future, imperative, conditional'.

5.6.4 Issue 2: Skills that may not be assessable

The assessment design proposed in this thesis sacrifices nonverbal communication. This is however not the only limitation; there are further ones the effects of which may not be as apparent. Below, several examples will be presented of hindrances to judgment that can be perceived from the current perspective. These examples do not refer only to skills, but also to situations (the assessment of skills to master these situation may be hindered) and person-groups (the assessment of their skills may be affected).

Case in which assessment may be not possible	Reason
Situations in which communication does not address a limited number of foreseeable topics but jumps form topic to topic (for example, references to arbitrary current events)	The method of modelling social situations permits only talking about elements of social situations visible on the screen (Section 4.3).
Skills for which aesthetic aspects matter: an enjoyable or motivational speech, a romantic conversation, etc.	The number of sentence stubs is limited and must suit many different stories (Section 5.6.3). Thus, the choice of sentence stubs cannot include variants featuring fine nuances of expression.
Behaviour under emotional attachment (e.g. firing an employee)	The stories are very short (Section 4.8.2). Thus, even by using video sequences, it is not likely to be possible to build up any emotional attachment to any agent.
Nonverbal social skills, nonverbal aspects of other social skills, and skills that relate to intense emotions	The user has no means of nonverbal expression (and hence limited means to express intense emotion). Furthermore, there are only limited means to perceive nonverbal expressions of the NPC (e.g. an NPC can smile).
People who have no general computer skills	The validation study excluded elder participants

	that may have not grown up with computers (Section 5.2.3)
People who do not have a certain minimum required level of abstract thinking or skills to handle the necessary cognitive load	Subjects in the validation study were students and professionals of various sorts (Section 5.2.3). Though the results show no correlation with cognitive load for this sample, this does not mean that no such skills are required, since the game requires abstraction from real social situations.

Table 28: Situations, skills and groups of people that that may not be assessable

5.6.5 Summary and Conclusion

In sum, the authoring of assessment stories appears to be a learnable task involving a substantial but acceptable amount of effort. The technical problems of extension of the method to further stories appear to be resolvable. However, these statements cannot yet be made on the basis of empirical evidence. For that, further research will be needed. The issue that several skills probably cannot be assessed is likely not to be resolvable, as the hindrances often stem from basic properties of the method such as the limited choice of sentence stubs.
Which remaining skills, then, are foreseeably assessable? For the PM Game, there are essentially two criteria that must be fulfilled:

- The interface must facilitate that the user shows the presence or absence of these skills in his or her game behaviour.
- It must be possible to identify the behavioural markers making up a skill using logical expressions referring to elements of the log files (Section 4.7).

In view of this and of the skill areas that may have to be excluded, as named in Section 5.6.4, the following can be reasonably assumed:

Further research is needed to determine to which skills the game will be extensible. From current point of view, it can be assumed that the verbal aspects of skills to master work-related social situations will be assessable if they involve interaction within a limited range of formalized behaviors.

This claim can be further underpinned by the following work in related areas:

Originally, Dede (1995, p. 49) formulated the above claim regarding the application domain of an e-learning simulation by Stevens (1989): ‘While the application built by this project focused on code inspection as the skill to be trained, through similar means preparation could be provided for a wide variety of work-related situations that involve social interaction within a limited range of formalized behaviors’. This e-learning solution relates to the previous interface most similar to the PM Game. See Section 5.1.3 for further details and screenshots of the interface used.

In a study discussed in detail in Section 5.5, Funke & Schuler (1998) assessed the skills needed in the following situations on the basis of written responses; all are examples of the skill area outlined in the box above:

- Customer courtesy in a routine situation
- Customer wishes conflict with regulations
- Customer with risky financing needs

- Subordinate must convince superior to delegate a consulting task to the subordinate
- Subordinate must inform his/her superior about an error
- Superior must justify higher-order organizational goal requirements to subordinates

This is the previous result most similar to the assessment method proposed in this thesis. Though it is not a computer-based method but uses human assessors, it equally relies on judging social skills from identifying predefined contents in written text responses.

6 Summary and Conclusions

6.1 Summary

A computer-based method for the assessment of social skill has been proposed. Since 1927, there has been documented research on assessment methods such as multiple-choice and later computer-based methods for automatising the assessment of social skills. So far, no such method has been successfully validated. The method proposed in this thesis has been validated for three different social skills in the area of leadership and customer communication.

6.2 Discussion

When a claim is made and substantiated, the reader may expect a discussion on what exactly has been resolved, what has not, and what is the novelty of the contribution, what is not. In order to perform this discussion in detail, its contents have been moved to separate sections in the evaluation in Chapter 5:

- How does the method relate to previous methods?
  The answer contained in Section 5.1 can be summarised as follows: while elements of the method have been known, the combination is new.
- Is it a solution for three isolated social skills or has the general problem of social skill assessment been resolved? The answer as a result of the discussion in Section 5.6 is: further research is needed to determine to which skills the game will be extensible. From the current point of view, it can be assumed that verbal aspects of skills to master work-related social situations will be assessable if these involve interaction within a limited range of formalised behaviours.

6.3 Limitations and further directions of research

The goal of this thesis has been the proof of concept that it is possible to assess social skills with this method. There are limitations, which urge further research:

- The assessment by the proposed computer-based method covers only the verbal aspects of the skills. Based on the previous result cited in Section 5.5, it is evident that verbal communication is in fact responsible for part of the variance of social skills, but certainly not for all. As pointed out in Section 1.1, it appears to be reasonable that this limitation can be covered by methods such as a short face-to-face interview or role-play, designed for the assessment of nonverbal skills. A next logical step would be to carry out a study to investigate this. The question is essentially how situationally specific nonverbal skills are. The author believes that much of the situational variance of nonverbal communication skills relevant for a job or learning goal could be covered by a small number of test situations. That is, while the breadth of situations relevant for a job or a learning goal could be addressed by a verbal-only testing method as proposed here, a short test for nonverbal communication could suffice so that in total these add up to a complete assessment of social skills. Such a separation of assessing the nonverbal and the verbal skills may actually even increase the overall quality of social skills assessment by paying due attention to both of these different aspects.
- Extensibility is an important criterion for a technical assessment method (Sections 2.4 and 5.1.3). However, extensibility of the PM Game has been supported only by arguments and by a qualitative study of letting persons develop game stories (Section 5.6). Certainty should be achieved by actual validation of game stories to assess further skills.

- Discriminant validity (Campbell & Fiske, 1959) has not been established for the method in general: is the game actually able to differentiate sharply between the assessed skills or is much of the variance of the assessment scores unspecific? Being free from typical human assessment issues such as the halo effect (Section 3.2.2), the proposed computer-based method may be attributed a chance to exceed human assessment. Results presented in Section 5.3.8 indicate that the assessment by the game might have discriminant validity. However, this can be properly examined only if a multitude of skills are assessed.
- Empirical validation studies should generally be repeated. In the particular case of a skill assessment method, on which potentially the job chances or educational advances of people could depend, repeated evidence of validity should be considered a must of responsible research.
- An additional test could be run to exclude correlations of the social skill scores with various sorts of intelligence test scores (compare Section 5.3.7).
- Questions towards a possible future industry application have to be addressed, such as for example, the following:
    - Efficiency of the assessment method (amount of assessment information generated per unit of user time).
    - Social validity: do users accept it as a testing method for social skills?
    - Adverse impact: do certain genders or ethnic groups have disadvantages when taking this test?
    - Efficient authoring leading to valid stories: when an industrial assessment centre is developed, it is not validated to predict job performance. The latter is a very labourious process that has been carried out in research. From this, research guidelines have been developed for assessment centres. The same should be achieved for the PM Game. This will enable creating game stories for industrial use without the labourious effort of validating each story.

6.4 The results in the terms of design science

This thesis commenced by citing the criteria of design science research in Section 1.3. Hence, this section will return to this topic and investigate if and in which way these criteria have been met by this thesis. The central criteria of design science are: relevance and addition to the ‘knowledge base’. The practical application potential (i.e. relevance) has been presented in Chapter 2. The more interesting question, however, is the fulfilment of the criterion to produce a contribution to the knowledge base. The proposed computer-based assessment method is an addition, which previously did not exist. Furthermore, the effect of valid social skills assessment has not yet been achieved using computer-based methods. Does this sum up all the new contributions? The technical means needed for the present dissertation were available already in 1990: programming languages and databases, graphics, albeit on somewhat crude screens, and the computer mouse. No other means are needed to program a PM Game. Is this true? The understanding of the sociotechnical system of social skills assessment was not fully available but was rather continuously developing. This comprises the test taker, the test, the interactions between them and last but not least, the employers and educators. Insights about this sociotechnical system have been developing for a long time: from the advent of social multiple-choice tests in the first half of the twentieth century, the increasing use of tests in employment decisions, the many attempts to resolve the difficult problem of natural language understanding by computers, research on assessment centre validity, the inflated expectations towards e-learning at the beginning of this century, until today, when the use of online recruitment and testing is becoming commonplace in

practice and well-founded in research. Insights about the sociotechnical system of social skills assessment have been discussed in the different chapters of this thesis and are summarised as follows:

- Accumulated evidence, that multiple-choice interfaces (which present the test taker with alternatives to chose from) do not permit the assessment of social skills (Section 3.1.1)
- An increased awareness for authorability: it is not enough to produce a novel solution for a few content examples; rather, contents must be changeable and extensible, otherwise they would likely end up unused as the previous solution most similar to the PM Game (Section 5.1.3). Authoring shall be feasible both in terms of effort as well in terms of skills demanded from the authors. Without possibilities to extend contents, any assessment solution is of limited practical use because large test batteries are needed to cover many skills and to provide several tests for the same skills (Sections 2.2 and 5.6)
- Understanding of the problem of habitability of natural-language interfaces (NLI) and the continued difficulties of natural-language understanding (Section 5.1.2)

These conditions ought to be considered when developing computer-based social skill assessment systems. There are reasons laid out in the referenced sections to support the opinion that implementation of valid computer-based social skills assessment is unlikely to be successful without complying with these requirements. The PM Game complies with these requirements as shown in Table 29 below:

Requirements	Particularly relevant aspects	Solution chosen in the PM Game
- Habitability	- Avoid natural-language understanding	- Menu-based communication
- Authorability	- Avoid 3D, photorealistic graphics, animated avatars, real time simulation - Avoid artificial intelligence	- Present social situations by means of text dialogues, icons and text descriptions - Generate NPC responses based on simple if-then rules using handcrafted text (‘infoojects’) instead of artificial intelligence - Reduce complexity of branching user actions and thus reduce authoring effort by keeping stories short, currently roughly around 20 user-NPC interactions - Ignore nonverbal communication
- Let the user invent possible responses in a social situation	- Avoid multiple-choice	- Sentence completion: the user forms complete sentences by filling in gaps of sentence stubs using icons that represent elements of the social situations - The set of available sentence stubs does not depend on the social situation

Table 29: Implementation of requirements for social skill assessment in the PM Game

To further substantiate the importance of the above three requirements, Table 30 presents examples for pieces of software that arguably failed by not adhering to one of these requirements:

Requirements	Unsuccessful attempts of social skill assessment violating the respective requirement
- Habitability (no natural-language understanding)	Paschall et al. (2005) – See Section 3.1.1
- Authorability (no AI, no state-of-the-art virtual world)	Stevens (1989) – This is not a social skills assessment system but an e-learning system. However, the system is in several regards similar to the PM Game and the argument of authorability is the same, see Section 5.1.3
- Let the user invent possible responses in a social situation (no multiple choice)	Donovan (1998) – This is not a piece of software but a multiple-choice test. However, multiple-choice tests are transferable into computer tests, thus considered equivalent in this thesis. Hence, the term 'automatic test'. See Section 3.1.1 for more examples.

Table 30: Unsuccessful attempts to assess social skills automatically

In sum, a contribution of this thesis contributes to the knowledge base is the collection of the above requirements and the arguments supporting the necessity of these requirements (at the current level of technological development) in order to build a computer-based social skills assessment system.

6.5 Conclusion regarding the core methodical elements responsible for the results

In conclusion of Section 6.4, the solution of combining menu-based communication with sentence completion can be considered as the core of the method responsible for the effect of valid social skills assessment. Clearly, it is in fact never one single methodical element that brings about such an effect in a practical implementation. Rather, many details are needed and even seemingly small usability issues may hinder successful assessment of social skills. However, the combination of menu-based communication with sentence completion is the methodical element that sets this approach apart from previous approaches attempting to assess social skills by the use of computers.

7 Appendix

7.1 Documentation of the PM Game stories and their assessment

The validation study (Section 5.2) was carried out using the following three PM Game stories:

- 'Hungary Hotel'
- 'The Problem'
- 'Mary and Garry'

Details of these stories and their assessment are described in the accompanying CD as well as in various sections of this thesis document in order to illustrate various methods and concepts discussed therein. Hence, this section serves as an index pointing to these distributed sources.

Details of the PM Game stories

'Hungary Hotel'

- Purpose: The story 'Hungary Hotel' served on one hand the purpose to exercise the game handling and hence to reduce noise data because of handling mistakes. On the other hand, pretests have shown that the first interaction with the game evokes substantial curiosity. In this regard the story had the purpose of 'venting' this curiosity. That is, the scores from 'The Problem' and 'Mary and Garry' are less noised by user actions of experimenting with the limits of the game.
- Storyline: Section 7.6.1
- Screenshots: several screenshots of the story are presented in Section 7.6.3.
- Example conversation: Section 7.4.1
- The complete set of all conversation logs recorded in the validation study is provided on the accompanying CD in the directory 'game log files'.

'The Problem'

- Purpose: assessment of the skill 'giving feedback to employee'
- Storyline: the user assumes the role of the manager of a construction company. The employee Luigi is the construction site manager, to whom the user delegates the task to build a copy of the Pisa tower. The customer is a rich person living in Ticino, the Italian part of Switzerland. While the copy of the tower is perfect, the process was not perfect: there have been issues of construction site safety. The situation is probably severe as the police were there and a representative of the workers reported issues. In this situation the manager has to give feedback to Luigi, who has in turn not been permitted to take holidays for a long time. The customer has ordered two copies of the Pisa tower. Talking to Luigi, the user must make sure that the issues of safety will be avoided when building the second tower.
- Screenshots: Section 4.3
- Example conversation: Section 7.4.3
- The complete set of all conversation logs recorded in the validation study is provided on the accompanying CD in the directory 'game log files'.

‘Mary and Garry’

- Pupose: assessment of the following skills
- ‘Eliciting customer requirements’
- ‘Establishing an emotional bond in a first customer meeting’
- while conducting a first meeting with a customer
- Storyline: the user assumes the role of the manager of an event management company ‘Live Your Dream Inc.’, specialised on outstanding weddings. The user has to conduct a first customer meeting with the couple Mary and Garry and find out about their requirements. They are to order a Hawaiian-style wedding.
- Screenshots: Section 4.5
- Example conversation: Section 7.4.2
- The complete set of all conversation logs recorded in the validation study is provided on the accompanying CD in the directory ‘game log files’.

Assessment of the PM Game stories

Due to its size, the complete documentation of the behavioural markers, scoring rules and participants scores of the played stories is provided on the accompanying CD in the following files:

- M&G - assessment by PM Game rules.xls
- TheProblem - assessment by PM Game rules.xls

An example of a scored log file from an actual play of a story is provided in Section 7.4.3.

7.2 Documentation of the role-plays

The following role-plays were used in the validation study (Section 5.2). They were developed by Ebert (2008) in order to evaluate the learning effect of the project management course of Stoyan (2008). In order to ensure accurate documentation of the study, the role-play scripts are presented in their original German version. English language descriptions of the storylines are presented below.

Role-play ‘Feedback Talk’

- Tested skill: giving feedback to an employee
- Storyline: the test taker assumes the role of the manager of a company in the educational sector. They apply for EU funding. For 1,5 months, the manager’s team has been working on writing the application. In this situation, the manager has to give feedback to an employee regarding his work and make sure that the application is successfully completed. The tasks of the employee has been to write the text for certain work packages that are part of the proposed EU project, to file the financial concept and to prepare the team meetings. The employee is behind the schedule with the work packages while the other work has been done to full satisfaction. To complete the application, there are 4 days left until the deadline.

‘Customer Meeting’

- Tested skills:
  - Eliciting customer requirements
  - Establishing emotional bond in a first customer converation
- Storyline: the EU proposal mentioned in the role-play ‘Feedback Talk’ has been submitted. A large bank is interested in sponsoring the same project. This sponsoring would constitute addi-

tional financial means for the project, which are highly welcome. The test taker is again in the role of the manager of the company. He or she has to conduct a first meeting with a representative of the large bank in order to find out about the conditions under which the bank would sponsor the project.

7.2.1 Role-play 'Feedback Talk' - participant information

ROLLENSPIEL
»RÜCKMELDEGESPRÄCH«

Instruktion für die Kandidaten

Ihre Firma, die d-a-q AG, ist zwei Jahre alt und im Bildungsbereich tätig. Sie sind Manager und Inhaber der Firma und auf der Suche nach zusätzlicher Finanzierung. Sie reichen einen umfangreichen Antrag auf Förderung des Bildungsprojektes „AppPermit" bei der EU ein. Dieses Bildungsprojekt planen Sie zusammen mit Partnern durchzuführen. Solch einen Antrag zu stellen ist jedoch ein Kraftakt, sie arbeiten bereits 1,5 Monate daran und stehen jetzt 4 Tage vor dem Abschluss. Ihr Team arbeitet mit Hochdruck an den zugewiesenen Aufgaben. Bei der letzten Arbeitsbesprechung ist Ihnen aber aufgefallen, dass Herr Trebe mit der Fertigstellung seiner Work Packages noch immer im Rückstand ist. Außerdem wirkt er leicht reizbar und unzufrieden. Beides zeigt sich insbesondere in den Arbeitsbesprechungen. Als Projektleiter kennen Sie seine Bedeutung für den erfolgreichen Abschluss des EU-Antrages und wissen, dass er aktuell noch zwei zusätzliche Aufgaben neben der Erstellung der Work Packages bearbeitet: einerseits kümmert er sich intensiv um das Finanzkonzept des Antrags, andererseits organisiert er die regelmäßigen virtuellen Arbeitsbesprechungen, ohne die ein systematischer Austausch und damit die Einhaltung der Deadline unmöglich wären. Beides nimmt zusätzliche Zeit in Anspruch, doch mit den Resultaten können Sie sehr zufrieden sein. Bei der Finanzplanung wurde der Spagat geschafft, die EU-Vorgaben einzuhalten, Einzelwünsche der Projektpartner zu berücksichtigen und ein ausgeglichenes Budget nachweisen zu können. Die Meetings sind gut vor- und nachbereitet, die Einladungen sowie Reminder werden rechtzeitig verschickt und die Ergebnisse transparent dokumentiert. Dennoch möchten Sie ihm eine Rückmeldung über Ihre Beobachtungen geben, weil Sie Herrn Trebe für den Endspurt noch einmal motivieren möchten und die termingerechte Fertigstellung des Antrags unter keinen Umständen gefährden dürfen.

Ihr Hauptziel für diese Aufgabe lautet daher:
Geben Sie Trebe Rückmeldungen zur bisherigen Leistung und finden Sie eine angemessene Lösung für eventuelle Probleme.

Sie haben nun noch ein paar Minuten Zeit, sich auf das Rückmeldegespräch vorzubereiten. Das folgende Rollenspiel »Rückmeldegespräch / Mitarbeiterproblem« dauert 10 Minuten. Bitte nutzen Sie die Rollenspielzeit sowie die Vorbereitungszeit aus. Viel Erfolg bei der anschließenden Aufgabe!

7.2.2 Role-play 'Feedback Talk' – actor's script

ROLLENSPIEL
»RÜCKMELDEGESPRÄCH«

Instruktion für die Rollenspieler

Sie sind Herr Trebe.
Das Projekt befindet sich kurz vor dem Abschluss, alle arbeiten mit Hochdruck an den ihnen zugewiesenen Aufgaben. Allerdings wurde bei der letzten Arbeitsbesprechung deutlich, dass Sie als Einziger mit der Bearbeitung Ihrer Work Packages im Rückstand sind. Sie wissen, dass diese Angelegenheit sehr wichtig ist und der Abgabetermin näher rückt. Allerdings bearbeiten Sie aktuell noch zwei zusätzliche Aufgaben im Rahmen des EU-Antrags neben der Erstellung der Work Packages: einerseits kümmern Sie sich intensiv um das Finanzkonzept des Antrags, andererseits organisieren Sie die regelmäßigen virtuellen Arbeitsbesprechungen. Beides nimmt zusätzliche Zeit in Anspruch, doch mit den Resultaten können Sie und Ihre Kollegen sehr zufrieden sein. Bei der Finanzplanung wurde der Spagat geschafft, die EU-Vorgaben einzuhalten, Einzelwünsche der Projektpartner zu berücksichtigen und ein ausgeglichenes Budget nachweisen zu können. Die Meetings sind gut vor- und nachbereitet, die Einladungen sowie Reminder werden rechtzeitig verschickt und die Ergebnisse transparent dokumentiert. Nun hat Sie Herr/Frau <NameTN> um ein Gespräch gebeten. Sie ahnen bereits, worum es geht. Sie sehen zwar ein, dass Sie im Verzug sind, aber Schuld daran tragen nicht Sie. Viel mehr haben Sie erst viel zu spät die Beiträge der einzelnen Projektpartner erhalten, deren Inhalte zu Work Packages zu verarbeiten sind. Die Regelung, dass Projektpartner nur von den Kontakthaltern kontaktiert werden dürfen, hat Sie bei der Arbeit behindert. So mussten Sie immer wieder zunächst bei den Kontakthaltern nachfragen und diese bitten, ihre Kontakte an die Frist und die noch ausstehenden Projektskizzen zu erinnern. Außerdem waren einige Skizzen so unverständlich formuliert, dass Sie erst nach mehrmaliger Rücksprache deren Inhalt in Erfahrung bringen konnten – und in einem Fall hat nicht einmal das zum Erfolg geführt.
Auch die Zusatzaufgaben haben zur Verspätung beigetragen. Die Erstellung des Finanzplans war ein Kraftakt, weil die Projektpartner sehr unterschiedliche Vorstellungen über ihren finanziellen Aufwand hatten. Deshalb mussten Sie viele Budgets kürzen, um einerseits die Antragsgrenze von 4 Mio. Franken nicht zu übersteigen und um andererseits keine allzu großen Budgetunterschiede zwischen den Einzelpartnern aufzuweisen. Außerdem scheint sich jeder das Recht heraus zu nehmen, an der Finanzplanung herum zu mäkeln und Änderungen zu fordern. Die Nachverhandlungen mit einzelnen unzufriedenen Projektpartnern hätten dazu geführt, dass Sie fast das Dreifache gegenüber der im Projektplan verankerten Zeit für die Finanzplanungsteilaufgabe benötigt haben. Vor diesem Hintergrund sind Sie sogar schon recht weit mit den Work Packages. Die Organisation der Arbeitsbesprechungen nimmt hingegen nur wenig Zeit in Anspruch, es wäre Ihnen jedoch geholfen, wenn jemand aus der Arbeitsgruppe die Dokumentation der Meetings übernähme.
Gerade in diesen stürmischen Zeiten hätten Sie sich mehr Führung und Unterstützung durch die Projektleitung gewünscht. Deshalb lassen Sie Kritik auch nicht auf sich sitzen, sondern reagieren mit Unverständnis. Sie haben bisher noch alles rechtzeitig geschafft und werden es auch diesmal schaffen. Auch die Vielzahl an Meetings nervt Sie zunehmend. Sie glauben, dass diese durch ineffiziente Moderation der Meetings durch die Projektleitung viel zu lange dauern und nahezu keinen Wert hätten. Solche Meetings lösen keine Probleme, sondern schaffen nur neue. Mit Ihrer Zeit wüssten Sie wahrlich Besseres anzufangen.

Details und Anmerkungen für den Rollenspieler:

Ziel des/der Teilnehmers/Teilnehmerin bei dieser Aufgabe ist es, Ihnen Rückmeldung zum bisher sehr positiven Projektverlauf zu geben und sich mit Ihnen über seine aktuelle Unzufriedenheit mit Ihnen zu unterhalten, weil er/sie in letzter Zeit Defizite in Ihren Arbeitsleistungen festgestellt hat. Nachdem er/sie Ihnen seine/ihre Beobachtungen und Bedenken beschrieben und Sie Ihr Problem geschildert haben, sollte der/die Teilnehmer(in) lösungsorientiert darauf eingehen und mit Ihnen gemeinsam Vorschläge erarbeiten, die Ihnen dabei helfen, Ihre Arbeit wieder in gewohnter Qualität ausüben zu können.
Sie sind zunächst defensiv, immerhin haben nicht Sie um das Gespräch gebeten, sondern Ihr(e) Chef(in). Soll der/diese also zunächst verraten, was er/sie von Ihnen will. Und das sollte zunächst Lob für die bisher hervorragende Arbeit sein. Sollte der/die Teilnehmer(in) darauf verzichten, können Sie ruhig beleidigt auf die vorgetragene Kritik reagieren. Viele Teilnehmer(innen) vergessen das Lob als Einstieg und scheuen sich davor, Kritik direkt anzusprechen. Wenn er/sie es also „von hinten durch die kalte Küche versucht", gehen Sie nicht darauf ein.
Wenn er/sie Sie auf seine Beobachtungen anspricht, ohne konkret zu werden, fragen Sie nach, warum er/sie frage oder was er/sie meine. Immerhin kann er/sie nicht erwarten, dass Sie ein Problem, das Sie bisher nicht angesprochen haben, plötzlich von sich aus erläutern, nur weil er/sie Sie nach Ihrem Befinden fragt. Sie haben schliesslich keine Zeit für ein Kaffeekränzchen und halten sowieso nicht viel von den Managementfähigkeiten Ihres Gegenüber. Wenn also jemand etwas von Ihnen möchte, dann soll er gefälligst den Anfang machen und direkt zur Sache kommen. Sobald Sie den Eindruck haben, unbestätigte subjektive Eindrücke bzw. Beobachtungen würden zu Tatsachen verklärt, Kritik pauschal und person- statt verhaltensbezogen geübt etc., verringern Sie ruhig Ihre Kooperationsbereitschaft. Kritisiert er/sie in angemessenem Maße, dann erklären Sie den Sachverhalt aus Ihrer Sicht. Erwarten Sie dazu Vorschläge vom Kandidaten. Geben Sie Herr/Frau <NameTN> bitte Raum, selbst Vorschläge zu entwickeln! Erst wenn das erfolgt ist, sollten Sie diese aufgreifen. Wenn dem/der Kandidaten/Kandidatin keine sinnvollen Lösungsvorschläge einfallen sollten, können Sie Ihre Forderungen anbringen. Dies könnte etwa mehr Unterstützung und effizientere Führung sein, zB in der internen und externen Kommunikation sowie bei der Moderation der Ihnen verhassten Meetings!
Darauf kann und sollte der/die Kandidat(in) zum aktuellen Zeitpunkt unbedingt eingehen, immerhin sind Sie der letzte mit ausstehenden Work Packages. Ergo sollten andere damit bereits fertig sein und über mehr Kapazität verfügen. Diese könnten Ihnen nun ohne weiteres helfen; bei welcher Aufgabe ist verhandelbar.
Ein Ziel des/der Kandidaten/Kandidatin besteht in der kollaborativen Entwicklung möglicher Lösungsstrategien. Lassen Sie sich daher nichts aufzwingen, sondern bestehen Sie darauf, bei der Lösungsentwicklung ein Wörtchen mit zu reden!
Es steht Ihnen natürlich frei, sich selbständig weitere Lösungsalternativen zusätzlich zu den hier notierten auszudenken.

7.2.3 Role-play ‘Customer Meeting’ - participant information

ROLLENSPIEL
»KUNDENGESPRÄCH«

Instruktion für die Kandidaten

Der Antrag auf Förderung des Bildungsprojektes „AppPermit“ wurde erfolgreich und gerade noch Termingerecht bei der EU eingereicht. Bei einem Abschlussgespräch mit einem Mitarbeiter, stellte sich heraus, dass dieser einen zusätzlichen potentiellen Geldgeber für das Projekt „AppPermit“ ausfindig gemacht hat. Jener wäre bereit, zusätzlich eine halbe Million CHF dem Bildungsprojekt beizusteuern, wenn gewisse Bedingungen erfüllt wären. Als Projektleiter sind Sie natürlich sehr an einem weiteren Geldgeber interessiert, da Sie auf Grund des knappen Budgets einige Aspekte des vollumfänglichen Projektes weglassen mussten, die jedoch sehr wünschenswert wären. Aus diesem Grund beschlossen Sie, den potentiellen Geldgeber (Herr Lender) zu einem Gespräch einzuladen und einmal seine Bedingungen zu prüfen. Auch wenn diese ihren Vorstellungen entsprechen würden, wollen Sie jedoch noch keine Verträge abschliessen, sondern am Abend das Ganze nochmals via Telefonkonferenz mit Ihren Mitarbeitern besprechen. Sie haben gerade noch Zeit Ihr Büro einigermassen besuchertauglich herzurichten und Ihr erst zweijähriges Firmenlogo „d-a-q AG“ von den daran geklebten Flipcharts zu befreien. Denn Herr Lender wurde schon vom Empfang gemeldet und ist auf dem Weg nach oben.

Ihr Hauptziel für diese Aufgabe lautet:
Führen Sie das Gespräch mit dem potentiellen Kunden und klären Sie seine Bedingungen.

Sie haben nun noch ein paar Minuten Zeit, sich auf das Kundengespräch vorzubereiten. Das folgende Rollenspiel »Kundengespräch« dauert ca. 10 Minuten. Bitte nutzen Sie die Rollenspielzeit sowie die Vorbereitungszeit aus. Vielen Erfolg bei der anschließenden Aufgabe!

7.2.4 Role-play ‘Customer Meeting’ – actor’s script

ROLLENSPIEL
»KUNDENGESPRÄCH«

Instruktion für die Rollenspieler

Sie sind Herr Lender.
Bei einer zufälligen Diskussion mit Herrn Müller, auch Mitglied in ihrem Ruderklub, sind Sie auf dessen Arbeitgeber „d-a-q AG“ und das Projekt „EU Antragsstellung »AppPermit«“ aufmerksam geworden. Sie arbeiten in dem europäischen Dienstleistungsunternehmen „EuroBank“ als Head HR Marketing und würden, falls der Antrag von der EU Kommission genemigt wird, gerne eine halbe Million CHF Budget beisteuern. Allerding haben Sie dafür auch einige Bedingungen: Das Hauptprodukt des Antrages – die Errichtung einer EU-zertifizierten Online-Bildungsplattform für Personen im Dienstleistungssektor – sollte nach der Erstellung von der Firma als „Imageverbesserer“ marketingtechnisch genutzt werden dürfen. Dies würde das Image von „Eurobank“ als Preferred Supplier im gewünschten Branchensektor nachhaltig steigern. Auch die Bekanntheit der Plattform

selbst würde durch diese Marketingaktion rasant zunehmen und „d-a-q AG“ würde somit doppelt profitieren.
Auf Grund dieses Doppelprofites sähe es die Firma EuroBank natürlich gerne, wenn sie für ihre Mitarbeitenden eine Ermässigung bei den Accounts zur Plattform erhalten würde.
Auch die Zahlung von einer halben Million CHF wäre eine einmalige Angelegenheit. Für die Folgejahre würde maximal ein Budget von ca. 30'000 CHF jährlich zur Verfügung stehen. Für Sie ist klar, dass dies nicht sehr gerne vom Gegenüber gehört wird und darum werden Sie diesen Aspekt nicht gleich erwähnen. Beim Nachfragen oder bei Fragen wie „sonst noch was“ etc. würden Sie jedoch dies schon bekannt geben.
An den Inhalt der Plattform stellt die Eurobank keine Forderungen, ausser das die gebotene Qualität des Bildungsstoffes erster Güte sein soll. Darüber machen Sie sich jedoch keine Sorgen, ist doch die „d-a-q AG“ ein renomiertes Team. Beim Nachfragen oder bei Fragen wie „sonst noch was“ etc. fällt Ihnen jedoch noch ein, dass die Bildungsplattform doch noch eine Unternehmensportrait-Seite der EuroBank beinhalten sollte.

Kurz zur Eurobank: Sie ist europaweit die drittgrösste Bank im Bereich der Vorsorge. Nur der Name ist noch unbekannt, da er auf Grund einer Fusion zweier Banken entstanden ist. Sie bieten ihren Kunden sowie ihren Mitarbeitenden Leistungen erster Güte.

7.3 Documentation of role-play assessment

The procedure of the role-play assessment in documented and confirmed below. The complete scores are provided on the accompanying CD and can be found in the files:

- ‘M&G - assessment by humans.xls’
- ‘TheProblem - assessment by humans.xls’

University of Zürich
Institute of Informatics

Binzmühlestr 14
CH-8050 Zurich
Switzerland
Tel. +41 76 3672572
stoyan@ifi.uzh.ch
www.ifi.uzh.ch/~stoyan

Robert Stoyan

Zürich, 18th of August, 2010

Confirmation of scoring system and non-influencing

Hereby I confirm that I assessed the mp3 recordings for the comparison study of the „PM Game" with two role play exercises. The study included 47 participants There were two assessors, me and Robert Stoyan. Before doing the assessments, a list of criteria and scoring guidelines was agreed, see the attached excel sheets. All assessments were done by each assessor independently by listening to the mp3 recordings of the role play exercises. While listening to mp3 recording, relevant aspects of the behavior of the participant were noted for each criterion. Based on these, each assessor gave a score according to scoring criteria and finally gave overall scores. There was no influencing of each other during this individual judgment process. Then the judgments of each assessor were compared and discussed only in cases where the noted observations differed or one assessor had the opinion that the other did not follow fully the scoring system. There was no discussion on the scores other than non-compliance with the scoring system. Finally, in case of the performance criteria, a separate mutual mark was decided for the main criteria and an overall mutual mark. The procedure for these mutual marks was free discussion in case of a difference.
During the whole listening and scoring process, both the participant´s real identity and their results in the PM Game were not known. I especially certify that I was not influenced in my judgments in any other way than stated above.
Each page of the attached final scores is signed by me.

Boris Otonicar
Lic. Phil. Psychologist

7.4 Example conversations

In order to illustrate the user–NPC dialogues in the PM Game, an example of a conversation log is presented for each of the three stories. They have been recorded as part of the validation study (Section 5.2) and contain the complete conversations that have taken place while a user played the respective story. In addition, the dialogues are used to present central concepts of the assessment method and the PM Game. These have been been introduced in Chapter 4.

Story	Demonstrated concept of the PM Game	Section in Chapter 4
Hungary Hotel (training)	–	–
Mary and Garry (assessment)	- Infoobjects - Speech act rules - Act with words	- 4.8.1 - 4.8.3 - 4.8.4
The problem (assessment)	- Behavioural markers - Skill scoring algorithm	- 4.6 - 4.7

Table 31: Concepts of the PM Game demonstrated in conversation logs

7.4.1 'Hungary Hotel' – story to learn the interface handling

The following conversation was recorded while a participant of the validation study played the story 'Hungary Hotel'. This story serves to learn the handling of the PM Game interface.

Conversation between the player and the NPC:
Player: Hi Tomi!
Tomi: Hi Dominic!
Player: Are you satisfied, Tomi?
Tomi: Yes.
Player: Please build the 1st leg!
Tomi: I did it.
Player: Thank you for having successfully built the 1st leg!
Tomi: Your are welcome.
Player: Thank you for having successfully built the 1st leg!
Tomi: Your are welcome.
Player: Thank you for having successfully built the 1st leg!
Tomi: Your are welcome.
Player: Please build the 2nd leg!
Tomi: I did it.
Player: Thank you for having successfully built the 2nd leg!
Tomi: Your are welcome.
Player: See you.
Tomi: Bye Dominic.
Player: Hello Mr Csödör!
Mr Csödör: Hello Dominic!
Player: I would like to present you the current project state:
Mr Csödör: Tell me more...
Player: I would like to present you the current project state:

Mr Csödör: Tell me more...
Player: Ok.
Player: Hello Mr Csödör!
Mr Csödör: Hello Dominic!
Player: Hello Mr Csödör!
Mr Csödör: Hello Dominic!
Player: I would like to present you the current project state:
Mr Csödör: Tell me more...
Player: I would like to present you the current project state:
Mr Csödör: Tell me more...
Player: the 1st leg is built.
Mr Csödör: I see
Player: Yes.
Player: I would like to present you the current project state:
Mr Csödör: Tell me more...
Player: I would like to present you the current project state:
Mr Csödör: Tell me more...
Player: the 2nd leg is built.
Mr Csödör: I see
Player: Are you satisfied, Mr Csödör?
Mr Csödör: Yes, I am.
Player: thank you for all you have done.
Mr Csödör: You are welcome!
Player: Good bye!
Mr Csödör: Goodbye Dominic.
Player: Hi Tomi!
Tomi: Hi Dominic!
Player: Please build the 3rd leg!
Tomi: I did it.
Player: thank you for all you have done.
Tomi: You are welcome!
Player: Yes
Player: See you.
Tomi: Bye Dominic.
Player: Hello Mr Csödör!
Mr Csödör: Hello Dominic!
Player: the 3rd leg is built.
Mr Csödör: I see
Player: Yes.
Player: I would like to present you our ideas:
Mr Csödör: Ok.
Player: Hello Mr Csödör!
Mr Csödör: Hello Dominic!
Player: the 3rd leg is built.

Mr Csödör: I see
Player: the bill is not paid.
Mr Csödör: I see
Player: Could you pay the bill, please!
Mr Csödör: Dominic, No chance to do that
Player: Hi Tomi!
Tomi: Hi Dominic!
Player: Please take apart the 3rd leg!
Tomi: I did it.

Table 32: Example conversation in the story 'Hungary Hotel'

7.4.2 'Mary and Garry' – assessment story

The following dialogue was recorded while a user played the story 'Mary and Garry'. In addition to showing how the story is played, the conversation also serves to demonstrate the application of the means of the game to produce human-like reactions. Reactions of the simulated interaction partners (NPCs) in the game are currently either produced by an 'infoobject' or by a 'speech act rule'. Infoobjects contain complex and highly situation specific reactions. Speech act rules contain simple 'general purpose' answers that are produced only when no infoobject is available. In addition, the conversation contains one example of an 'act with words' user action.

Conversation between the player and the NPC	Produced by
Player: Nice to meet you!	
Mary and Gary: Nice to meet you too!	Speech act rule
Player: My name is Filomena.	
Mary and Gary: My name is Mary and Gary.	Speech act rule
Player: Could you tell me more about the wedding?	
Mary and Gary: We would like a real Hawaiian style wedding. 100 persons will be invited.	Infoobject
Player: Tell me something about the decoration, please!	
Mary and Gary: The decoration should give an authentic Hawaiian flavor: palm leaves on the walls and lei for all ladies. Lei are traditional Hawaiian flower garlands.	Infoobject
Player: Could you tell me more about the dinner?	
Mary and Gary: The dinner... hm. Maybe "poi", thats a simple, but very old traditional Hawaiian dish.	Infoobject
Player: Ok.	
Player: Do you know something about the date?	
Mary and Gary: Thats a good question... Well, say in one year from now.	Infoobject
Player: What do you think about the budget?	
Mary and Gary: Yes, the budget. We discussed that and it should be around 100.000 Euros.	Infoobject
Player: I see.	Speech act rule
Player: May I summarize what we talked about?	
Mary and Gary: Please.	Speech act rule

Player: the date is in about one year.	
Mary and Gary: I see	Speech act rule
Player: the budget is 100.000 Euros.	
Mary and Gary: I see	Speech act rule
Player: the dinner is the Hawaiian national dish "Poi".	
Mary and Gary: I see	Speech act rule
Player: the decoration is lei for the ladies and palm leaves for the room.	
Mary and Gary: I see	Speech act rule
Player: the wedding is Hawaiian style.	
Mary and Gary: I see	Speech act rule
Player: Are you satisfied, Mary and Gary?	
Mary and Gary: Yes, I am.	Speech act rule
Player: I would like to present you our ideas:	
Mary and Gary: Ok.	Speech act rule
*** what you want to do *** i will prepare an offer and present it to you, ok? **********	Act with words (by the user)
Player: May I offer you a drink?	
Mary and Gary: Thank you!	Speech act rule
Player: Good bye!	
Mary and Gary: Goodbye Filomena.	Speech act rule

Table 33: Example conversation in the story 'Mary and Garry'

7.4.3 'The Problem' – assessment story

The following conversation was recorded while a user played the story 'The Problem'. In addition to showing how the story is played, the converation also serves to demonstrate the application of the scoring algorithm decribed in Section 4.7. However, for better understanding, not the original scoring rules are used here, but the substantially shorter, better readable behavioural markers. Excerpts of the scoring rules is provided where necessary to make the scores understandable.

The behavioural markers referenced below are taken from the complete matrix of behavioural markers, detailed scoring rules and participant scores. Due to its size, this matrix is provided on the accompanying CD, file name: 'TheProblem - assessment by PM Game rules.xls'. To each behavioural marker, there is an exact scoring rule. While the behavioural markers address contents (meaning of the conversations), the scoring rules operate on syntax. Thus, while the former are more suitable in this section for illustration purposes, the latter are substantially longer but contain all information needed for computer-based assessment.

Conversation between the player and the NPC	Scoring
Player: Hi Luigi!	
Luigi: Hi Filomena!	
Player: Please build the 1st Pisa Tower!	
Luigi: I did it.	
Player: Hi Luigi!	

Luigi: Hi Filomena!	
Player: We need to talk!	
Luigi: Ok.	
Player: Are you satisfied, Luigi?	
Luigi: Yes.	
Player: Tell me more about the safety!	The player failed the following behavioural marker (0 points out of 1): 'First, recognize positive aspects of the work done, in order to keep the motivation of the employee', because the player addresses the problem area 'safety' directly (it is known that there have been errors) instead of recognizing first the several months of successful work Luigi has completed. The scoring rule requires to not address the problem areas 'safety', 'Toni' and 'police', before providing any form of positive feedback. The player failed the following behavioural marker (0 points out of 1): 'If you know that critique is needed, tell the feedback yourself instead of asking what you know (pretending as if you would not know)' The player fulfilled the following behavioural marker (1 point out of 4): 'Talk about important problems'. Remark: the scoring rule accepts also a question or a request to do something as means to talk about a problem. There are 4 problems areas presented by the story that the player can talk about: 'safety', 'Toni', 'police', 'holidays'. The player fulfilled the following behavioural marker (1 point out of n): 'If you make a critical statement about something, give the employee the possibility to tell his or her own point of view.' Remark: the scoring rule prescribes that if the question comes first and the statement comes later, this also counts as fulfilment. 'n' is the number of

	things criticized. In this conversation log, n = 1.
Luigi: Oh... I made mistakes here! It was very hectic work. I forgot many details about the workers safety. I'm so overworked. Why don't you give me holidays?? I had no holidays all the year. May I just take 4 days off?	
Player: I see.	
Player: What you did led to the following consequences:	
Luigi: ?	
*** what you want to do ***	
the police was here	
**********	
Player: Please talk with the police!	The player fulfilled the following behavioural marker (1 point out of 4): 'Talk about important problems'. Remark: the scoring rule accepts also a question or a request to do something as means to talk about a problem. There are 4 problems areas presented by the story that the player can talk about: 'safety', 'Toni', 'police', 'holidays'.
Luigi: I did it.	
Player: Tell me something about Toni!	The player fulfilled the following behavioural marker (1 point out of 1): 'When talking about work the quality of which is not surely known, ask a question. Don't make a statement on its quality.' Remark: it is known that the workers representative Toni had some problem, but it is not clear what that exactly was. Hence, it is appropriate to ask a question instead of forming a judgment. The player fulfilled the following behavioural marker (1 point out of 4): 'Talk about important problems'. Remark: the scoring rule accepts also a question or a request to do something as means to talk about a problem. There are 4 problems areas presented by the story that the player can talk about: 'safety', 'Toni', 'police', 'holidays'.
Luigi: Oh, Toni is an attentive guy! I remember he noticed some safety defects, but nobody had the time to correct them. Maybe I should talk to him...?	

Player: Yes	
Player: Please talk with Toni!	
Luigi: I did it.	
Player: the safety is a little bit problematic.	The player failed the following behavioural marker (0 points out of 1): 'No belittlement', that is, serious problems should not be presented in a wording as if being harmless. The player fulfilled the following behavioural marker (1 point out of 1): 'Be specific when expressing critique' Remark: Though 'the safety is a little bit problematic' would not count as a specific critique in real role-plays, in the PM Game, this is the maximum specifity possible: at least the object of the problem has been mentioned correctl. The scoring rule specifies an exact list of sentences that count as unspecific critique such as for example 'I´m unsatisfied'. If only such very unspecific critique is said by the player, then the point for this behavioural marker is not avarded.
Luigi: I see	
Player: What do you think about the safety?	
Luigi: I promise, I will introduce sound safety measures: I will schedule a regular safety check on the building site, so that no one gets hurt.	
Player: Tell me something about the safety!	
Luigi: Thanks for asking once more. I see this point is important for you. I can assure you that I will completely solve this problem by scheduling regular safety checks and I have everything what I need to do so.	
Player: thank you.	
Luigi: You are welcome!	
Player: What do you think about holidays?	The player fulfilled the following behavioural marker (1 point out of 4): 'Talk about important problems'. Remark: the scoring rule accepts also a question or a request to do something as means to talk about a problem. There are 4 problems areas presented by the story that the player can talk about: 'safety',

	‘Toni’, ‘police’, ‘holidays’.
Luigi: I`m very satisfied with my job. But I would be so glad to receive some holidays. I just don't see the point in working all my life without a break. I don't know why you never give me holidays!	
Player: Tell me something about holidays!	
Luigi: I have no information	
Player: Do you know something about holidays?	
Luigi: I have no information	
Player: I feel glad that you have talked with Toni.	
Luigi: Oh	
Player: thank you.	
Luigi: You are welcome!	
Player: That was very good how you have talked with the police!	
Luigi: Yes	
Player: I feel satisfied that you have talked with the police.	
Luigi: Oh	
Player: Please take holidays!	
Luigi: I did it.	
Player: Please build the 2nd Pisa Tower!	
Luigi: I did it.	
Player: Tell me more about the safety check!	The player fulfilled the following behavioural marker (1 point out of 1): ‘Check if agreed improvements have been implemented’.
Luigi: Thanks for asking! I introduced regular safety checks during the construction of the second tower. Mamma mia! I was so surprised how much it increased the workers satisfaction!	
Player: Did you build the 2nd Pisa Tower?	
Luigi: Yes	
Player: You are doing a lot for us!	
Luigi: Oh	
Player: I am satisfied.	
Luigi: Ok	
	Total score: 4 points. Remark: the behavioural marker ‘talk about important problems’ has been completed 4 times, which was exactly the number of opportunities. Thus, it counts 1 = 4/4 points.

Table 34: Example conversation in the story 'The Problem'

7.5 Documentation of the sentence stubs in the communication menu

The following screenshots document the sentence stubs available in the communication menu. There are altogether 38 sentence stubs.

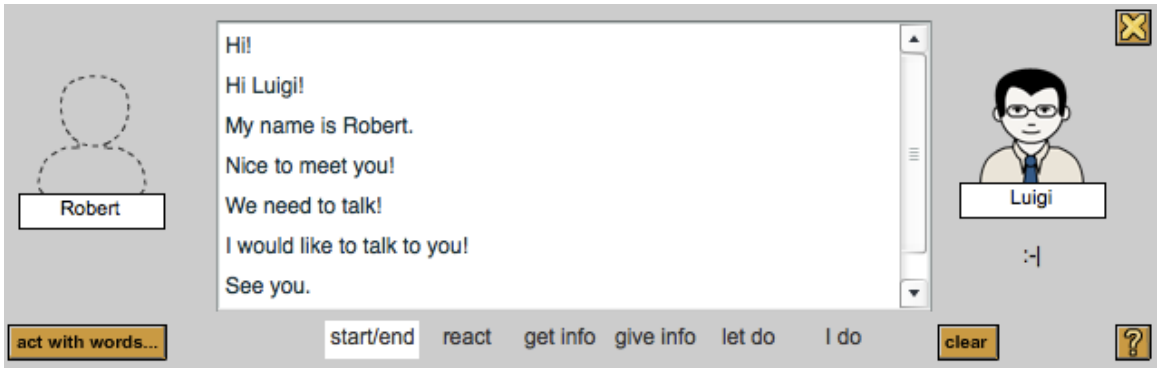


Figure 34: The ‘start/end’ menu: the first 7 sentence stubs

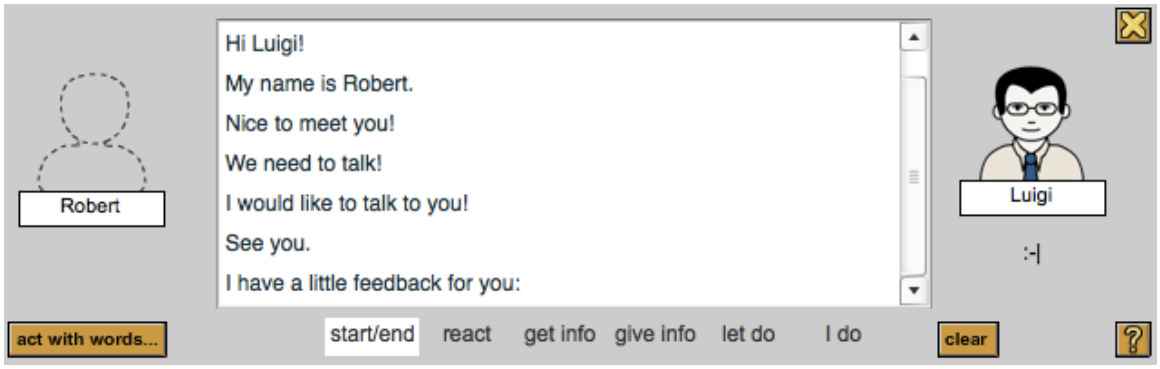


Figure 35: The ‘start/end’ menu: 1 further sentence stub

Figure 36: The ‘react’ menu: 6 sentence stubs

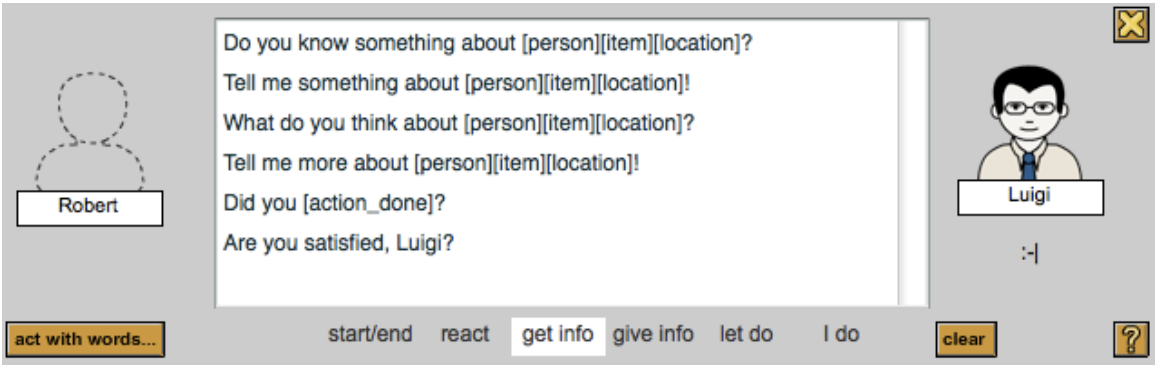


Figure 37: The ‘get info’ menu: 6 sentence stubs

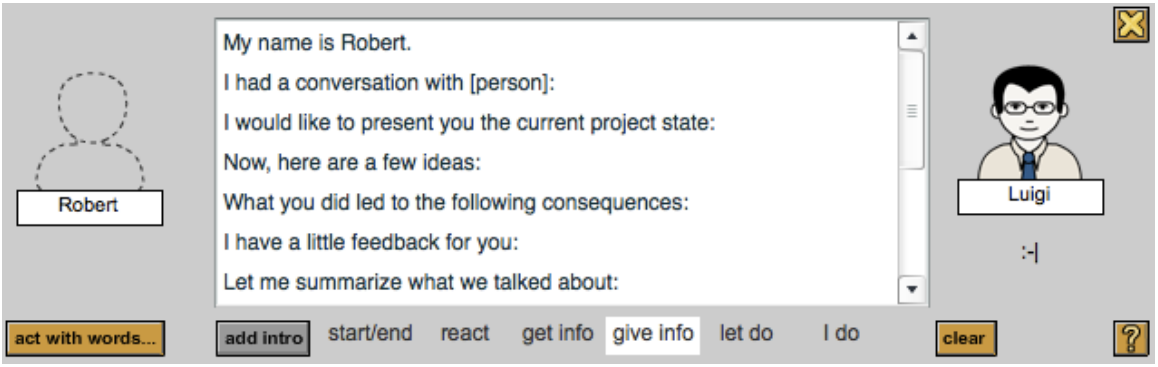


Figure 38: The ‘give info’ menu: the first 7 sentence stubs

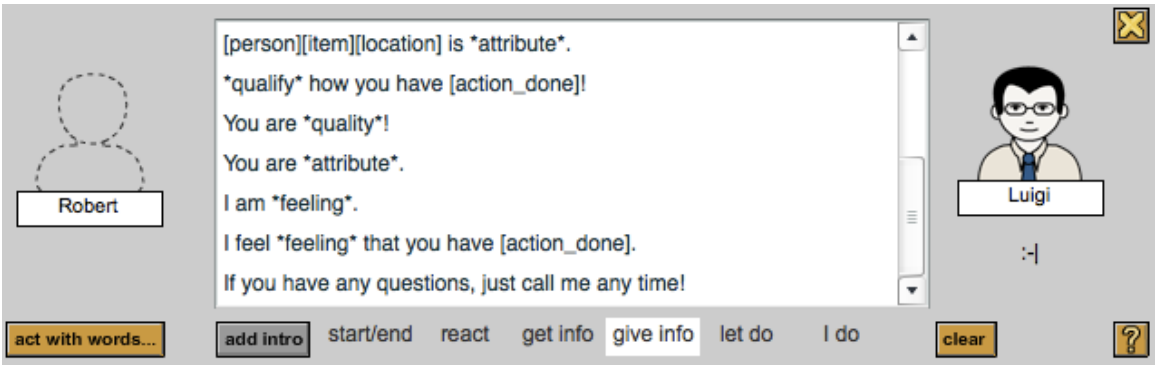


Figure 39: The ‘give info’ menu: further 7 sentence stubs

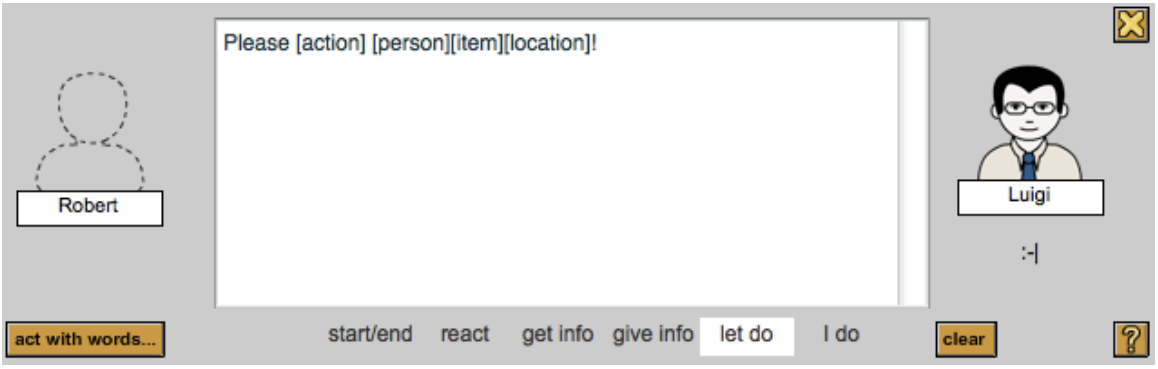


Figure 40: The ‘let do’ menu: 1 sentence stub

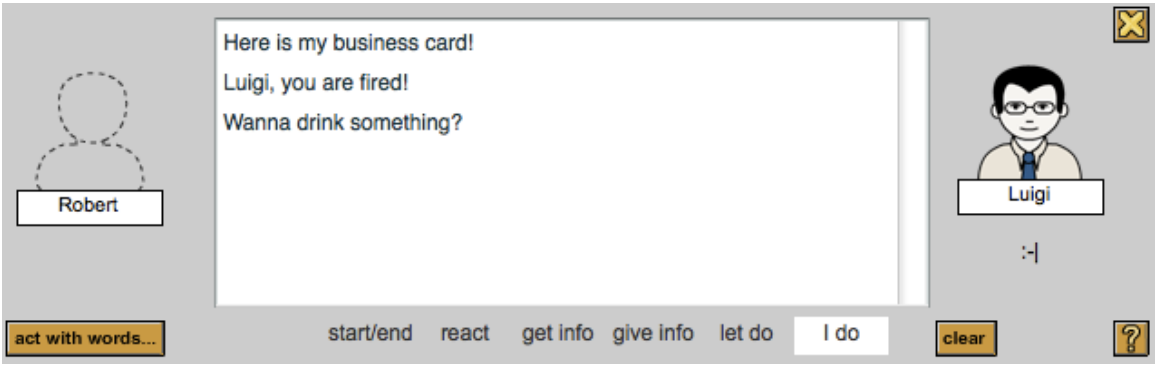


Figure 41: The ‘I do’ menu: 3 sentence stubs

7.6 Use of hidden goals to author the course of events in stories

In the PM Game, ‘hidden goals’ are the means used to model dependencies between events. For the author of a story, they are a general tool that can be used to influence the course of events. The concept is described in Section 4.8.5. Using the story ‘Hungary Hotel’ from the validation study (Section 5.2) as an example, this appendix illustrates how hidden goals function. In this section, first, the story is explained. Then it will be shown how the goals are defined in the editors (author’s view). Finally, the sequential events in the story are presented when they are triggered by the goals (user’s interface).

7.6.1 Hungary Hotel

Storyline: in ‘Hungary Hotel’, a customer is unwilling to pay for the work done by a construction company. As a somewhat radical solution, the user (in the role of the manager of the company) has to order the removal of a valuable part of the house (the golden cupola). This forces the customer to pay up.

Background: an interview with a Hungarian construction site manager revealed that removing valuable parts from a building is actually a tactic used to reduce the loss of the construction company when the customer is unwilling to pay. This often forces the customer to pay up.

Purpose: this story was played by the participants of the validation study as a training in order to learn how to handle the game interface.

7.6.2 Use of goals to trigger events

In the story sketched above, a goal is used to trigger certain events. The user’s action of removing the golden cupola fulfils the goal ‘demolish the 3rd leg’. This causes the customer to pay the bill and

makes the customer's facial expression sad. Furthermore, it causes a notice to appear that the user has successfully finished the story. Figure 42 to Figure 45 show the editing process.

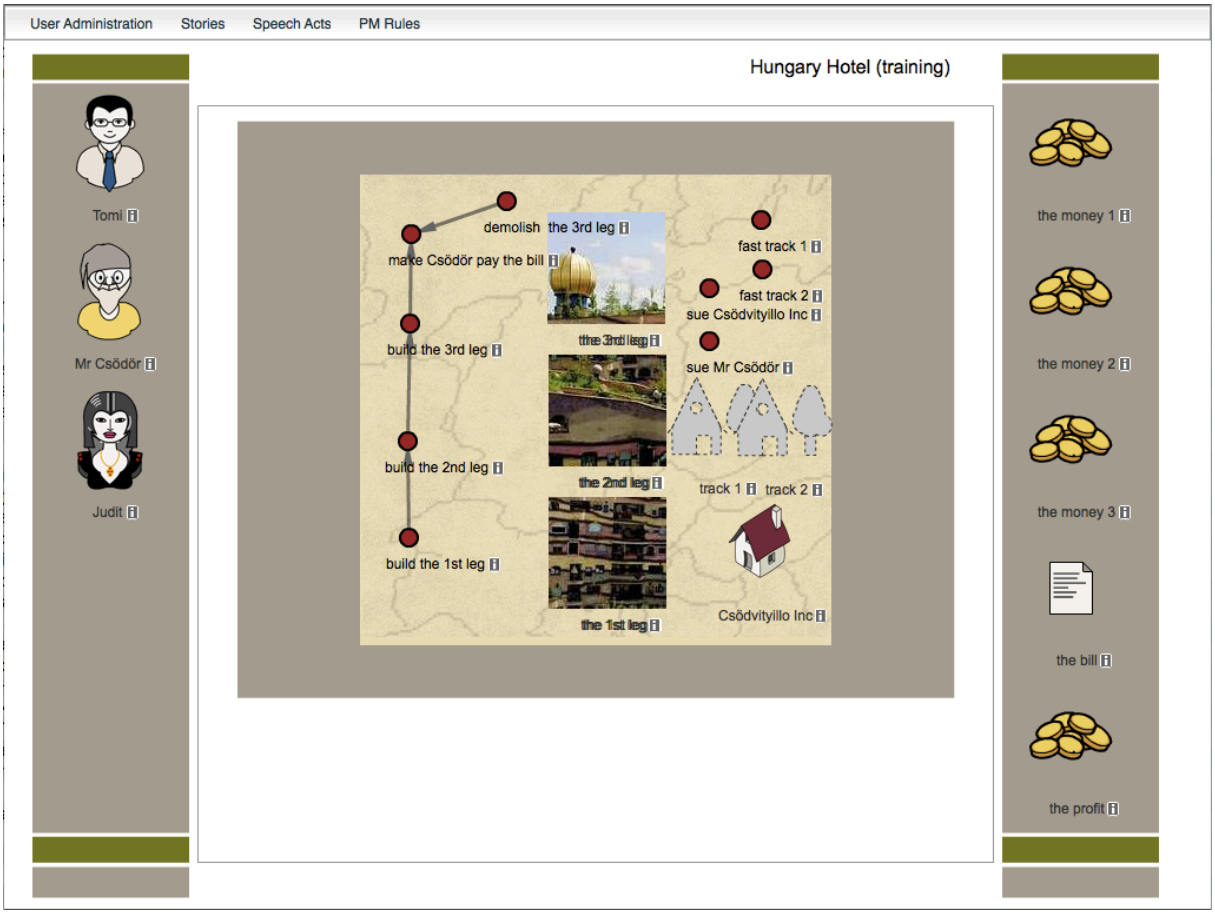


Figure 42: Main window of the game story editor showing 'Hungary Hotel'

There are nine red circles in Figure 42 that denote the goals. These goals are hidden; that is, they only appear in the game editor but are not visible in the playable story.

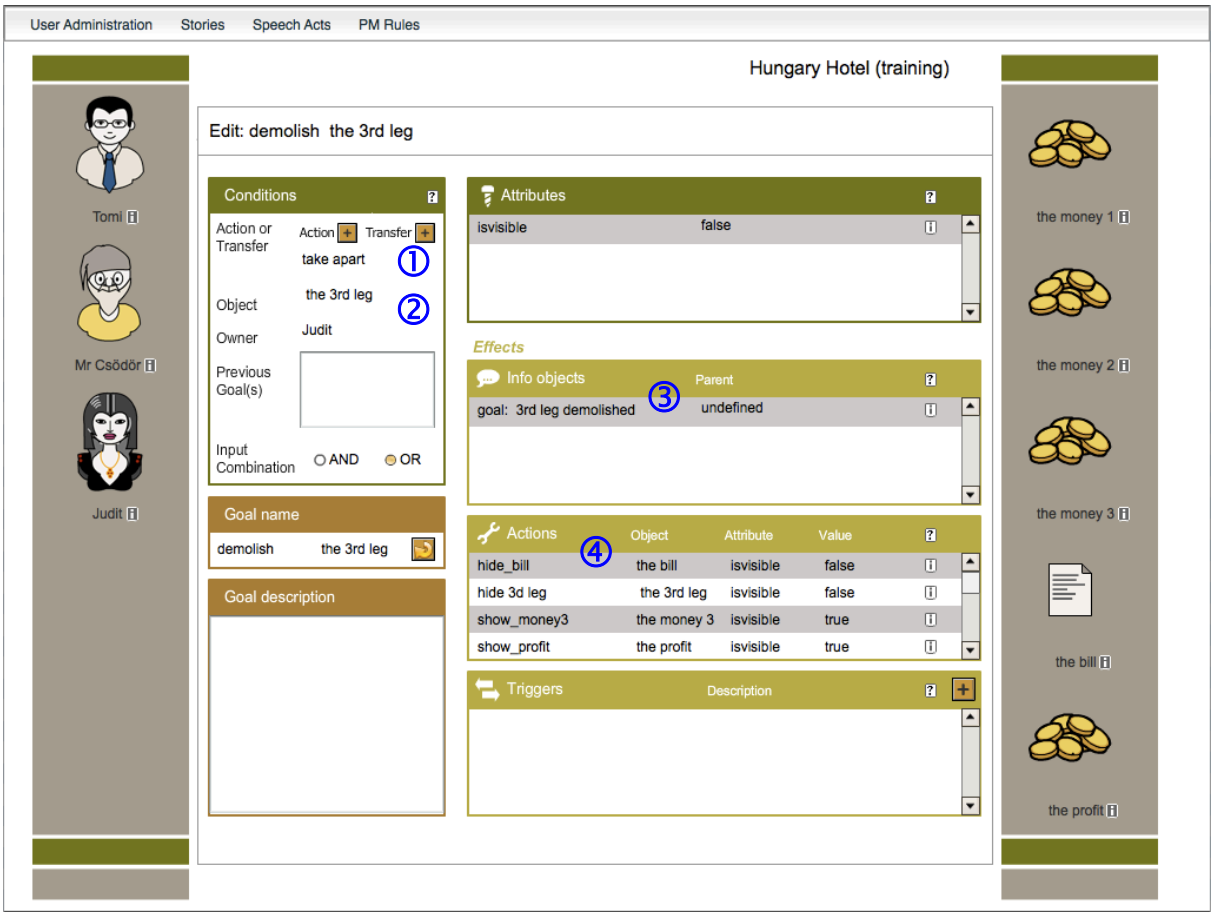


Figure 43: Editor window for the goal 'demolish the 3rd leg'

The 'conditions' subwindow in the upper-left corner of Figure 43 shows how the goal is fulfilled: the action 'take apart' (1) needs to be applied to the object '3rd leg' (2). This then leads to a cascade of events, such as the display of a message with the ID '3rd leg demolished' (3). The message contains a notification for the user that the customer, Mr. Csödör, has paid the bill and that the user has successfully completed the story. Further events induced are listed in the subwindow 'actions' (4) at the bottom, including the hiding of the bill and 3rd leg and the showing of various money-related items. Scrolling down in the 'actions' subwindow reveals further events, such as the emotional changes of the NPCs:

Actions	Object	Attribute	Value
Judit_smile	Judit	display_rule	:-)
csödör_angry	Mr Csödör	display_rule	:-(
show_money1	the money 1	isvisible	true
show_money2	the money 2	isvisible	true

Figure 44: Further actions initiated by fulfilment of the goal 'demolish the 3rd leg'

Taking apart the golden cupola is the end of the story. Prior to this, the user is guided to request the customer to pay the bill, but the customer refuses. This process is modelled by adding another goal: 'make Csödör pay the bill'.

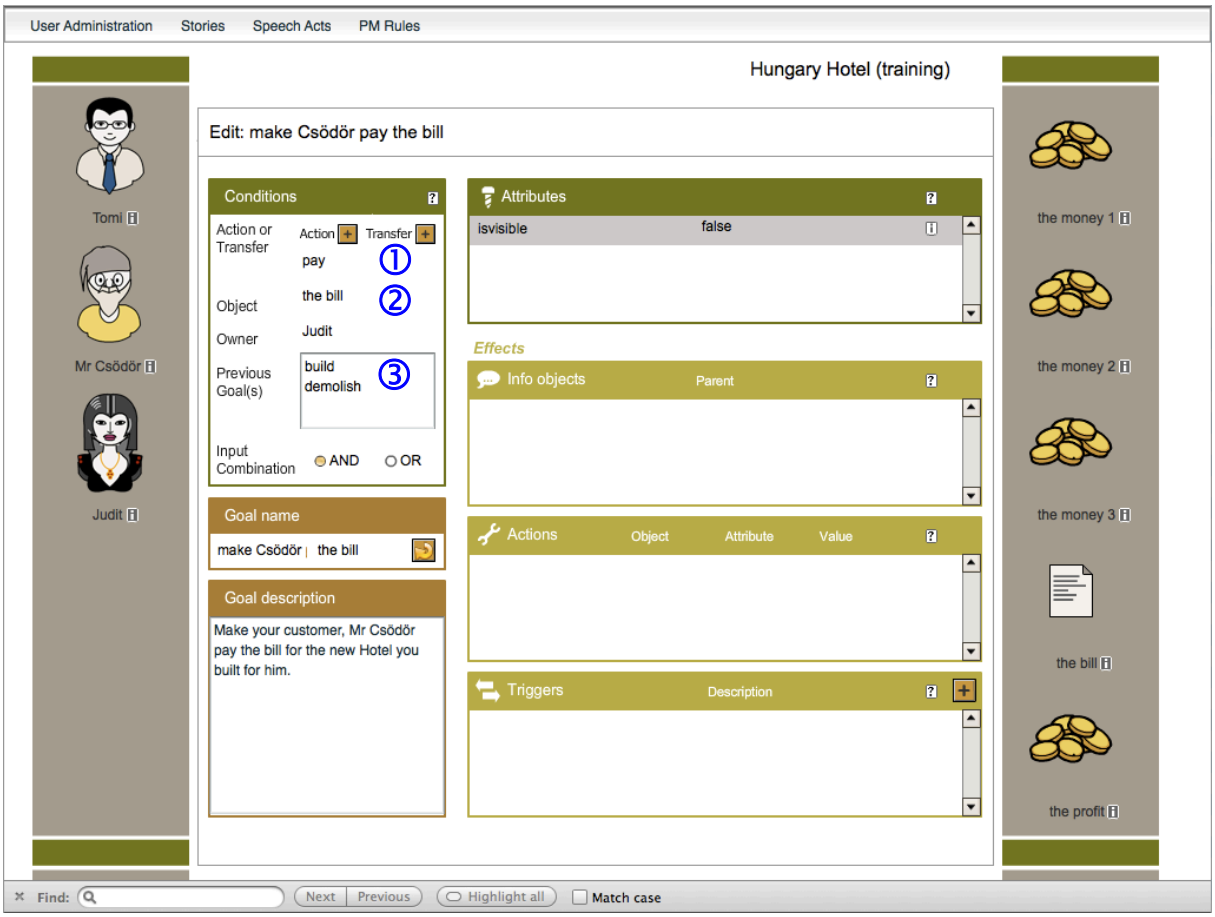


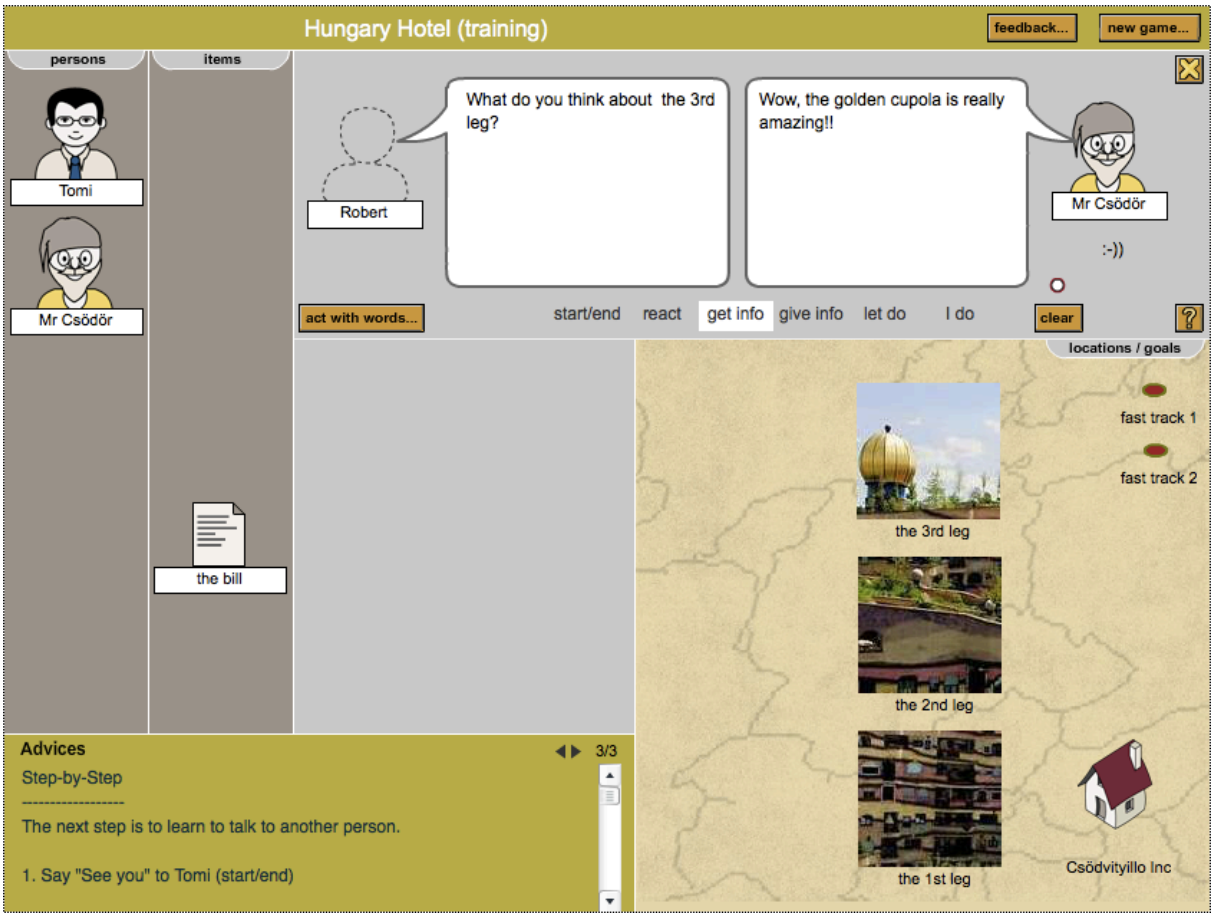
Figure 45: PM Game editor showing the goal ‘make Csödör pay the bill’

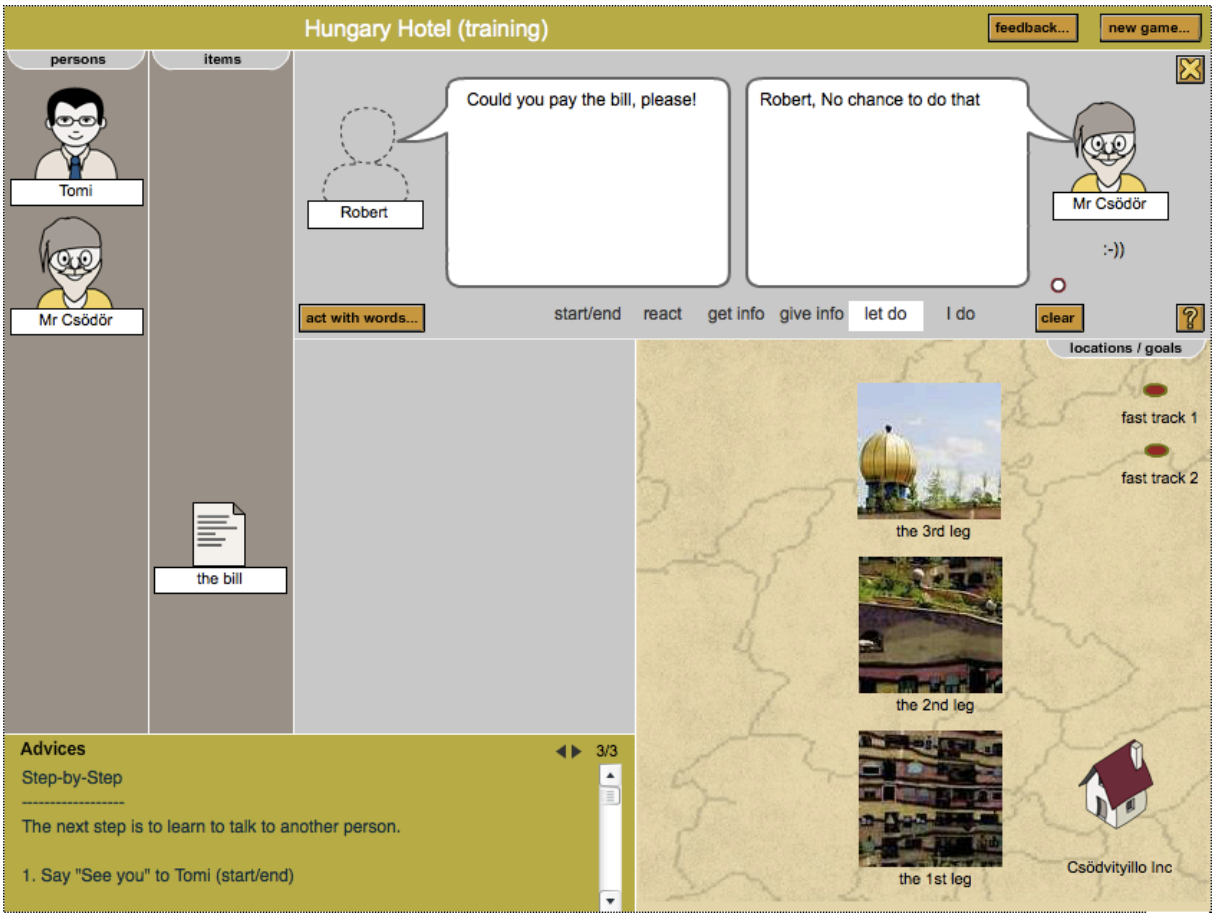
As shown in Figure 45 above, the goal ‘make Csödör pay the bill’ is fulfilled when the action ‘pay’ (1) is applied to the object ‘bill’ (2). However, there are two preconditions (3), which are comprised of the previous goals: build (the 3rd leg) and demolish (the 3rd leg). The action of payment can only be executed when these preconditions are fulfilled. That is, this goal acts by withholding the action of payment and delays it until the preconditions are fulfilled. This leads to the desired course of the story: the 3rd leg (the golden cupola) first has to be built; Mr. Csödör has to be asked to pay the bill; the execution of this is delayed because the precondition of taking apart the 3rd leg is not fulfilled; and when the 3rd leg (the golden cupola) is taken apart, the goals are fulfilled, and the delayed action is executed.

7.6.3 Resulting story

In the played story, the effect of the configuration of goals explained above is shown in the following sequence of screenshots. The first few screenshots display the preceding events. As shown in the screenshots, the user does not personally execute any of the actions (such as taking apart the 3rd leg). As it is a communication game, the user delegates the actions to his or her employee, who executes them.


Hungary Hotel (training)
feedback...
new game...
persons
items
Tomi
Mr Csödör
Robert
Please build the 3rd leg!
I did it.
Tomi
:-)
act with words...
start/end
react
get info
give info
let do
I do
clear
Information
The 3rd leg has been built successfully!
Now it is your task in this story, to finish the project by making your customer, Mr Csödör, pay the bill. Complete this on your own!
the bill
locations / goals
fast track 1
fast track 2
the 3rd leg
the 2nd leg
the 1st leg
Csödvityillo Inc
Advices
3/3
Step-by-Step
The next step is to learn to talk to another person.
1. Say "See you" to Tomi (start/end)


Hungary Hotel (training)
feedback...
new game...
persons
items
Tomi
Mr Csödör
Robert
What do you think about the 3rd leg?
Wow, the golden cupola is really amazing!!
Mr Csödör
:-))
act with words...
start/end
react
get info
give info
let do
I do
clear
locations / goals
fast track 1
fast track 2
the 3rd leg
the 2nd leg
the 1st leg
Csödvityillo Inc
the bill
Advices
3/3
Step-by-Step
The next step is to learn to talk to another person.
1. Say "See you" to Tomi (start/end)


Hungary Hotel (training)
feedback...
new game...
persons
items
Tomi
Mr Csödör
the bill
Robert
Could you pay the bill, please!
Robert, No chance to do that
Mr Csödör
:-))
act with words...
start/end
react
get info
give info
let do
I do
clear
locations / goals
fast track 1
fast track 2
the 3rd leg
the 2nd leg
the 1st leg
Csödvityillo Inc
Advices
3/3
Step-by-Step
The next step is to learn to talk to another person.
1. Say "See you" to Tomi (start/end)

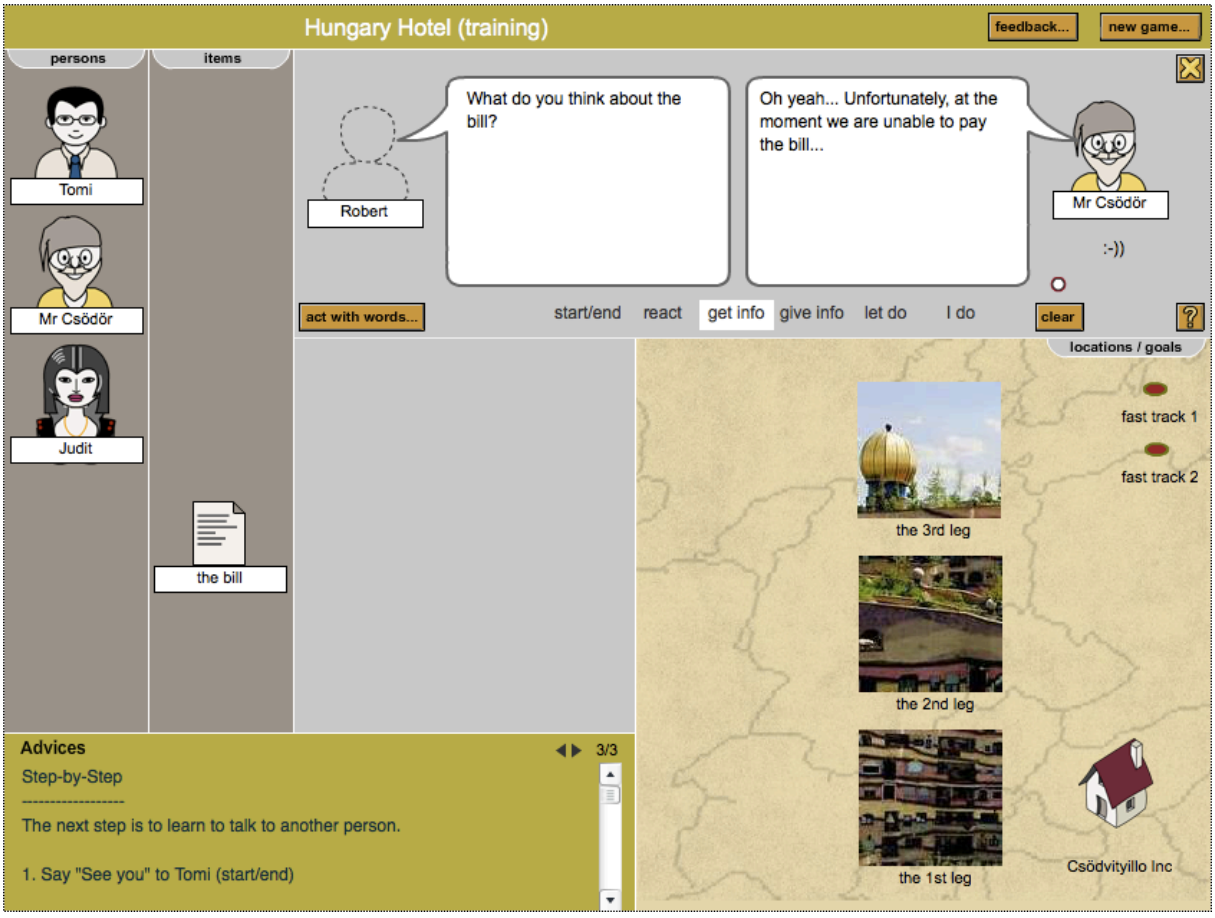

Hungary Hotel (training)
feedback...
new game...
persons
items
Tomi
Mr Csödör
Judit
the bill
Robert
What do you think about the bill?
Oh yeah... Unfortunately, at the moment we are unable to pay the bill...
Mr Csödör
:-))
act with words...
start/end
react
get info
give info
let do
I do
clear
locations / goals
fast track 1
fast track 2
the 3rd leg
the 2nd leg
the 1st leg
Csödvityillo Inc
Advices
3/3
Step-by-Step
The next step is to learn to talk to another person.
1. Say "See you" to Tomi (start/end)

Figure 46: Tomi has built the 3rd leg

Figure 47: Mr. Csödör is delighted to know about the golden cupola

Figure 48: Csödör refuses to pay the bill

Figure 49: Any further negotiation with the smiling Mr. Csödör is hopeless

For the sake of brevity, a few additional steps in the story have been left out (consulting Judit, the lawyer of the construction company). Finally, the drastic solution is displayed in Figure 50.

Figure 50: Solution

As shown in Figure 50, the hidden goals are fulfilled, and thus a cascade of events occurs: the disappearance of the cupola, the face of Mr. Csödör turning sad, Mr. Csödör paying the bill, and the display of the notice of success to the user.

7.7 Guidelines for the author of PM Game stories

The following guidelines were compiled from the experience of the supervision of several persons while they were creating stories and the personal experience of crafting the stories presented in this thesis. These guidelines are presented as part of this thesis to document in order to make authoring a repeatable process. Furthermore, the guidelines can be read as a list of the major problems around

which stories have to be built (limited pool of user sentences) and the major pitfalls (see hints for the various steps below) that were encountered in the past and thus should be avoided.

Guidelines

Introduction

It is easy to create stories for fun in the PM Game. Authoring useful stories is a professional endeavour that needs some creativity and, primarily, a suitable method. The goal is to create a story in the PM Game that is easily playable by the people in a target group and that tests a specific social skill. PM Game stories test social skills, such as how feedback should be given to an employee or how a relationship with a customer should be developed. The following is a hands-on training guide that explains the authoring process step by step.

Target group

This document is written for subject matter experts who are not software engineers but have expertise in their own domains: for example, leadership, customer communication or intercultural communication.

Overview

The total duration required to create the first playable story that tests a small skill is approximately thirty full working days; this task is distributed over a time span of at least two months. The estimated duration includes the time required for the creation of the story, learning the usage of the game editors, and, in particular, learning the authoring process. The estimate does not include the time required for the validation of the story. The time duration strongly depends on the availability of users to test the game at various steps of the story creation process as well as the availability of literature, experts, or personal experiences to define the test content.

Major work phases:

- Step 0: Learn to use the story authoring technology.
- Steps 1–7: Learn the methods required to create a story around certain behavioural markers (knowledge of what type of behaviour typically leads to success or to failure in a situation).
- Step 8: Repeat the previous work of creating a story by including more behavioural markers in order that the story can cover a small skill area.
- Step 9: Develop rules to score user performance.
- Step 10: Validate the story, or prove that the story indeed assesses the skill under consideration. This last step is not fully described in this guideline and is a task that takes several months.

Step 0: Create a simple story in order to learn how to use the game editors

Time: Half a day. Requisite: editor access to the PM Game
A guideline for using the game editors to create simple stories can be found here:

http://www.pm-game.ch/wiki/images/2/2a/Pmgame_tutorial.pdf
Experiment with the game editors for a few hours, as this is necessary for understanding the remaining content of this guide.

Step 1: For your first meaningful story, start with the behavioural markers (expert knowledge) you want to hide in it instead of directly starting with the story!

Time: Try to complete this in 15 min or less. Requisite: text editor such as Word

The main tool for assessing a certain social skill using the PM Game is to 'hide' behavioural markers in stories. An example of a behavioural marker is the wisdom of providing a summary at the end of a customer meeting. In general, a behavioural marker is defined as behaviour that is a sign of competence or incompetence in a defined social situation. 'Hiding' it in the story refers to the process of crafting a situation in which this behaviour is important for producing a successful outcome even though there is no obvious hint indicating that such a behaviour should be used.

Unlike the approach used in a scientific process, do not first read literature or conduct interviews with experts to find behavioural markers because not all the brilliant insights that you discover will be possible to implement in the PM Game. (Remark: Once you are done with your first story, then read literature or interview experts to find behavioural markers. When doing so, you will greatly benefit from having already created your first story as you will be able to recognize usable behavioural markers.)

Concrete steps

- Chose a situation for your story. Your story will test or teach the user how to handle this situation. The situation definition can be very short; for example, 'give feedback to an employee'. It is strongly recommended that you be familiar with the situation from academic and/or practical experience. Logically, you should not try to test people on a topic you are not familiar with yourself. If you have a good reason to still do so, do not do it for your first story because the necessary preparation (reading, talking to experts) will take weeks. Most of that time will be useless because you will not know which information from your reading is useful for the game.
- Start with only two to four simple, general, and useful behavioural markers. These can be anything plausible for the situation you want to test. An important note: all these behavioural markers must be applicable in the same situation.

Try to complete Step 1 in about 15 min. There is no need to try to be perfect because the game technology contains many surprises, and you first need to learn those before you can define perfect behavioural markers.

Step 2: Build a dialogue from the available sentence stubs!

Time: 2–3 hours. Requisites: text editor and the PM Game

When playing the PM Game, you will notice that it has a large but fixed set of sentence stubs: for example, 'Hello!' or 'What do you think about …?'. They form the core of the game. It is very unwise to create stories that need additional sentences. A single new sentence can change the entire game and affect preexisting stories. Often, this will also mean substantial programming effort. Thus, refrain from doing this and use the existing sentence stubs.

Reason: The problem is that sentence stubs in the PM Game are global; that is, each sentence stub is available for all stories. This means that all the old stories become unplayable if a single new sentence is added, as there are no prepared answers to this new sentence in those stories!

Detailed explanation: The sentence stubs are intentionally made global. If sentences differ from story to story, users will browse the sentence list to infer the actions possible in specific stories. For a proper assessment, it is very important that the game does not give clues! If you give clues, such as the availability of specific sentences, then it becomes clear to the user that there must be something important about these new sentences in the story.

Consequence: Do not start by envisioning a nice story as if you were writing a book. This will result in the need for many sentences that are not available in the PM Game.

Instead, follow the steps given below:

- Create your story sentence by sentence from the available sentences! Have your computer running, the PM Game open, and always check, sentence by sentence, what the next sentence of your story should be. What are your options when designing NPC answers? As you have learnt in Step 0 of these guidelines, an NPC is a person steered by the game, and they have two types of answers to a sentence selected by the user: handcrafted (‘infoobjects’) or automatic. Thus, the NPC in your dialogue must answer a user sentence with either of these types. To determine the NPC answers to a user sentence, just use the sentence while playing an existing game story. Answers that are simple such as ‘I see’ or ‘Hello’ are usually automatic answers. More complex ones are infoobject answers. Now, repeat the sentence a few times. First, the infoobjects are presented (if there are any); then, if they have all been used, the same automatic answer is repeated indefinitely. You can create the infoobject answers yourself, but you cannot change the automatic answers. You can only overwrite them with infoobjects, because the latter have higher priority.
- Compose a dialogue from the available user sentences and automatic answers, adding handcrafted answers (infoobjects) where needed.
- Write this dialogue in a text editor, such as Word.
- Create a short dialogue—maybe five sentences or so in total–and try to involve one to three behavioural markers from the few that you collected! What does ‘involve’ mean in concrete terms? Example: if the behavioural marker is ‘summaries are a sign of competence in customer meetings’, then all you need to do is to create a story where the user meets a customer and the story is crafted in a way that it makes it very likely that several pieces of information are talked about. If several pieces of information are given in a customer meeting, then this provides an opportunity for the user to summarize. In short: your story should give opportunities for the user to show a behaviour defined by your behavioural markers.

The following advice has proven to be useful for the people who have created stores until now. These are presented here as part of Step 2 because your text dialogue may already contain a few issues. After reading the advice, and doing the exercises contained herein, check your dialogue for the described issues and modify if needed.

→ Minimize the use of abstract objects/persons!

Sometimes, it is important to talk about abstract objects such as a risk. Try this task now: create an object with the name ‘the risk’ and choose a symbol for it, e.g. a bomb. The user should be able to talk

about ‘the risk’. Once you have done this successfully, try something more complex: create a ‘placeholder’ object/person like ‘the man who brought the fruit’ and then play your story and enunciate ‘Peter is the man who brought the fruit’; this will cause issues. Try this now to experience the limitations of the game firsthand. (One issue may be that the name is too long. Also, when playing the story, having weird context-less communication fragments in the game is awkward.)

→ Avoid explicit choices!

An ‘explicit choice’ means a situation where the user is presented with a small number of options such as ‘you can either do A or B’. Some persons who have authored stories so far have devoted substantial efforts into creating beautiful stories presenting explicit choices to the user before realizing that for such an explicit test there is no need to use a complex process such as the PM Game. You can easily write a multiple-choice test instead in a text editor in a few minutes. The problem with these tests is that they test only academic knowledge or possibly test intelligence or problem-solving abilities. There is a long scientific explanation for why multiple-choice tests cannot test social skills; the short version is that in real-life social situations, nobody tells you what the alternatives are! To elaborate, social skills (including intercultural skills) are not simply knowledge but behavioural patterns. Such patterns are (mostly unconscious) pieces of knowledge that are activated according to certain clues contained in a social situation. Explicit options do not activate behavioural patterns in a similar manner as a social situation, in which there are usually no explicit choices. Explicit choices activate rational/analytical thinking; that is, test takers start making comparisons and draw logical conclusions, such as when solving school exercises, and that tests their knowledge, intelligence, or problem-solving skills. In the very best case, you will have created a test of social sensitivity (ability to recognize critical points in a situation), but this is not a social skill. A counterargument could be that the game is artificial anyway and appears on a computer screen, which are not normally involved in social situations. This is undoubtedly correct and indeed provides clues that distort the perception of the game as a social situation. However, human beings have imagination; we do not live in the real world but in the one our mind creates from what is seen. Thus, if you have ever felt fear when reading a vampire story, then you know how imagined stories can induce real fears. The PM Game works similarly.

The above text describes what not to do. What should you do then? Use implicit choices! Again, think of the example of summarizing at the end of a customer meeting. In a customer meeting, there is an implicit choice that one can do this, but nobody tells you that this is an option. Clearly, if a user tries all of the many options provided by the game, then he or she will be confronted with this option to summarize in the form of the sentence, ‘May I summarize what we talked about?’ However, since there are so many sentences, it is not really an option presented like in a multiple-choice test. It is simply inefficient for the user to browse through all of the many options and compare which are the best. It is more efficient for the user to consider the best way to react and then try to map this to the game. In this way, a test for social skills can be created!

Step 3. Create variants of your dialogue

Time: 1–2 hours. Requisites: text editor and PM Game

What happens to your story if the user chooses different sentences then those you planned? Write down variants of your dialogue. Most importantly, write down what happens if the user violates your behavioural markers. Such an action by the player will sometimes change the dialogue completely and sometimes not at all. For example, if the story is about communicating with a customer and the user

does not provide a summary, then the meeting will end as usual. However, in other cases there are consequences: e.g. the customer becomes dissatisfied. As the story is for assessment, rather than causing consequences that should be avoided, they may provide the user with a clue as to what he or she did wrong. Even worse, the user may start looking for such clues.

Step 4. Play against a real human with a paper prototype

Time: 2 hours including preparations plus a few hours to modify your story. Requisites: 1–2 other people and a piece of paper (or Skype).

Now, test your story for the first time with a real human; you still should not have implemented anything with the PM Game. This usually makes sense as implementing a story in the game requires substantial work and is often in vain because it does not work with other people. To prevent this from happening after the implementation, gain some experience by testing the story with other people. There are two successful versions of this process so far:

a) If you have somebody who is familiar with the PM Game, ask this person to play the user. You play the computer. In other words, his or her task is to produce PM Game style sentences as if he was playing your story and using the PM Game interface. This can be done using Skype or any other Internet chat system.
b) This version can be played with anyone:
   Give the person a piece of paper. Write on it:

   begin/end react give information get information let do sthg do sthg

   Giving more details such as the sentence lists in the PM Game has not been shown to provide any benefits. Now, verbally explain the situation with which your story starts (this is the replacement for the written story introduction in the PM Game). You then ask the person to always point to one of the above options and then verbalise what he wants to do. For example, point to the begin/end and then say: ‘Hello’.

Both these tests are obviously not an accurate imitation of what happens in the PM Game, but you can gain several insights:

- What actions people are likely to do in the situation you present. Hence, you gain insight as to whether the available user sentences are sufficient. You also get insights into what NPC answers need to be produced
- Optional: if you know the person very well, then you will receive some insight into whether the test you created is a true test. You will see if the behaviour in the game largely corresponds or does not correspond to the typical behaviour of the person in real life. This is an interesting option that has been successfully used in the past; however, more experience is needed to determine how to use this and how much to expect.
- If you have several people you play with, you can then try to determine if your behavioural markers perform well. A simple criterion is that there must be people who produce and who do not produce your behavioural markers. For more on this little piece of test theory, see Step 7. If a behavioural marker is not produced by anybody or produced by all of the people, exchange it now for another one. A test is a means to uncover differences among people.

Step 5. Prepare for anything to be said at any moment!

Time: 2–4 hours. However, this step can easily extend to a full day and involve the rewriting of your entire story. Requisites: text editor and PM Game

Normally, authoring stories (as in books, theatre or movies) involves linear thinking: the medium dictates what comes after what. Digital storytelling is completely different from this. The beauty and main advantage of digital stories is that the user can say anything at any time. Thus, check your story step by step: what would be the reaction if the user chooses to say any of the many possible sentence combinations? This is a substantial amount of work. It usually results in the realization that the story can easily be ruined by the user: the user will say a sentence that is meaningful but that was not planned for in that particular situation, and the NPC answers with something meaningless. An example of this may be an infoobject of the customer saying how much he likes a new product before he had any chance to learn about the product. Though surprises like this are nearly unavoidable, there are methods of sequencing stories to some extent once they are discovered. The freedom of the user to say anything at any time remains, but pieces of information that must come later indeed do:

- Infoobjects can make objects visible. Try the following: create an object and then set it as invisible. Try it out and start playing your story; you will not see the object. Now, attach the object to an infoobject. Play the story again, make the infoobject be said, and the object appears! This often helps meaningless NPC reactions to be avoided for the simple reason that as long as the object is not visible, the user cannot click on it to talk about it.
- Infoobjects may have predecessors. Even if you say or ask something that would normally trigger the infoobject, it will not appear until the predecessor has already been triggered. This is commonly used to determine which reaction is to be said first, second, etc. Advanced use: the predecessor system also works across different trigger objects (e.g. ‘What do you think about x?’ and ‘What do you think about y?’).
- You can use goals. Goals are hidden objects that you see in editing mode on the map. With goals, you can model practically any kind of sequencing you want: for example, make an object disappear when some condition is fulfilled. This can be very useful if you do not want the user to talk about an object anymore! Tricky usage: have two objects with the same name. With this you can make your NPC talk differently about a house which is being planned and another which is already built. One of the objects can be named ‘house’, and the other one can be ‘house ’. No one will notice the difference. You can attach infoobjects about the planned house to one of these objects and the infoobjects about the ready house to the other one. The usage in user sentences will be the same, but the NPC reactions will be different!

Step 6. Implement your first small story in the PM Game

Time: minimum 1 day. Requisites: PM Game, your previously created texts, and several test persons

Hopefully, you kept to the rule of putting only a few behavioural markers into your story. This will save you substantial time. Now, create the story in the game. However, do not think that your story is finished when you enter into the editors of the PM Game. Test it! Even if you try it on yourself and it works, it is still not finished. It is only complete when other users have played it, you have corrected the observed errors, several users have played it, you observed them playing, and it proved to work. This is why you need so much time for this step. Thus, do this now: after implementing and testing your story, test it with a few other people, correct any errors, test with other people again, etc. Why

were these three steps listed in a single step in this guide? This is because in software development (and creating digital stories is software development), implementing, testing, and improving always belong together on principle. A piece of software that is only implemented is only one-third ready. You may notice that testing and especially reworking things so that they really work is almost as much work as all the work you have done so far in total. You are now a software developer!

Step 7. Check if the story indeed tests what you wanted it to test

Time: minimum 5 hours, up to 1–2 days if you need to rewrite your story. Requisites: PM Game and several participants

At this point, you have tested if your story is playable with users and probably reworked it a few times. However, do you know if it actually tests what you want? Determining this is a more laborious endeavour. You need several users who you let play your game. At this stage, it must be users who never played your story before. You must observe them without helping or interacting with them at all. The safest method of doing this is to not sit at their side and to not observe them. Observing users was useful for creating a running story. For this step, it is better to check only the resulting game protocols. You will see what they did in the end anyway!

There is a method to foresee if your story is a good test of a social skill: some of the users should fail, and some should do it right. For example, some users should forget the summary at the end of a customer meeting and some should remember. The obvious reason for this criterion is that a skill test is to differentiate people who have a skill from people who do not. Thus, a test is a means of finding differences among people. What if your story does not fulfil this criterion? A reason may be that you did not have the right test persons. Define the target group of your test! Invite people who are representative of your target group. If you indeed have the right people belonging to your target group and the test did not have both players who failed and succeeded, then the number of people may have been too small. It is ok if a single behavioural marker does not have equal numbers of failures and successes. The division may be skewed as much as 10% one way and 90% the other way, and it is still a good test. However, if nobody fails or nobody succeeds, then it is not a test for this target group. Hence, your test needs to be improved by changing your story. Also, you may not be able to improve your test because what you want to test is very well known or not known at all by all of the people in your target group. In that case, you need to abandon that behavioural marker.

Step 8: Create a story testing one small skill

Time: about 10–15 days. Requisites: text editor, PM Game, around 10–15 persons of the target group, and literature, experts, or personal practical experience with regard to the task and situation for which the story is crafted

So far, you have crafted a story in Steps 1–7 around a few behavioural markers. However, test efficiency is very low if people play one story (usually around 10–15 minutes) just to determine if they show a few behavioural markers or not. Thus, try to create a story (or two stories closely related to each other, that is, relying on the same context) that tests a small skill. The point is that when playing a story, much time is spent on understanding the context: that is, reading the introductory text and examining the available objects and NPCs. This time should be leveraged better by testing for many behavioural markers.

Concretely, do the following:

- Start the entire procedure again from Step 1, but this time do the following as well:
- Chose a skill with a size similar to the following examples of small skills: ‘giving feedback to an employee who has performed a job delegated by you that is partially good and partially wrong’. To cover bigger skill areas such as ‘leadership’ with one or two stories is obviously not possible. Define your skill as the ability to master a task (in the example above, the task is to give feedback). It should be a task typical and relevant for a certain job (in this example, leading a team). Specify a typical and relevant situation in which the task is to be carried out.
- Cover this skill with about a dozen behavioural markers. For this, it is a good idea to collect a pool of around 20 behavioural markers from literature, expert interviews, or personal experience. You need this many because a number of them will not be implementable in the PM Game or just not produced by people playing the PM Game, etc.
- The behavioural markers must cover the complete skill. A further explanation is as follows. For example, if the skill is ‘building up an emotional bond with the customer’, it is a good idea to reduce this to the specific situation of the behaviour in the first meeting and to the case of preparing a long-term customer relationship. However, the skill defined in this way needs to be covered fully by behavioural markers. In other words, there should be a behavioural marker with regard to most typical aspects that need to be covered in such a conversation to make it a success. Thus, if only the beginning or end of the conversation, attention to the customer, or the appropriate presentation of oneself are tested, then either the skill definition needs to be reduced or the number of behavioural markers increased. This aspect is important because otherwise the test is wrong: it claims to assess a skill but assesses only a small part of it.
- Keep the story short. The minimum number of user turns (how many times the user says something) required to successfully master the story should not be substantially more than the number of behavioural markers tested by the story. Longer stories cause a list of problems; most of all, the history of events leading up to the production of a behavioural marker cannot be controlled very well. The longer the story, the more often a situation that produces a behavioural marker will not be reached at all; instead, the user will do something unplanned so that the story ends up somewhere else. Thus, the test will be unable to determine if the user has the skill or not. The simplest but best trick to writing effective assessment stories is to directly present the user with a situation where the behavioural markers can be produced and then end the story.
- Minimize cognitive load. This means that you should not present the user with a complicated story because you are then testing his or her ability to understand complicated stories. Equally, do not present much information at once as you are then testing the user’s ability to handle a large amount of information at the same time. If the presentation of the story cannot be simplified any further but is still too complicated, then you may present all required information in a phased manner: first present a situation and then ask the user to do an exactly specified action, which will lead to a change in the situation that brings about the rest of the information.
- You need to rigorously test the story with other people and then improve it. Around 10–15 persons from your target group will be needed until you determine that the story works.

Theory: There are different possible definitions of a ‘skill’. The old method of traditional assessment centre psychology was to define so-called ‘dimensions’ as skills. Dimensions are generalized aspects of behaviour that can be observed across many situations. An example of such a dimension is the abil-

ity of a person to push through his or her opinion. This can be observed in many different situations, such as a group meeting with customers or a one-on-one meeting with a single employee. This theory was proved to be wrong. More exactly, such skill aspects exist, but they are weak. In other words, when somebody has a strong ability in one situation to push through his or her opinion, it is not necessarily true that this holds also for other situations. Behaviour is strongly dependent on the situation. The newer way to define a 'unit of skill' is to simply consider the ability to successfully complete a task as a skill. This also a better fit for the PM Game, in which the main unit of organisation is the playable story, which is a situation in which a task is to be carried out by the user.

Step 9: Determine scoring rules

Time: about 4 weeks of elapsed time, which works out to about 10 days of full-time work. Requisites: 20 persons of the target group, PM Game, SPSS or any software that can calculate correlations.

After completing Step 8, your story is finished: it is playable and you see that the behavioural markers work (some of the people fail and some succeed). The task in Step 9 now is to formalize the behavioural markers as computer rules (what exactly counts as right or wrong) and to make a first statistical analysis as to whether these formal rules are suitable for a skill assessment.

Concretely, do this:

1. Analyse previous game protocols and write down a scoring rule for each behavioural marker. The simplest scoring rule is as follows: if the behavioural marker is done, score +1; if not, score 0. More complicated cases are as follows:
2. Sometimes there are preconditions; for example, a behavioural marker may only be meaningful when a specific thing has happened previously. (For example, scoring the quality of feedback for a delegated task requires the delegation to have happened before. However, depending on the story, it may be possible that the user has not delegated anything.)
3. Observe what the successful and less successful users do. From this, conclude the range of expectations. A typical example of a behavioural marker that needs a range of expectations is 'active listening' (indicate by comments such as 'hmm', 'yes', 'indeed', or 'interesting' that you are following what the other person says). Here, the maximum expectation in a specific story was that such a behaviour is done twice (some users do it more often, but that is rare), a medium performance is doing it once, and a poor performance is not doing it. These performances are scored 1, 0.5, and 0 points, respectively.
4. Recruit 20 persons of your target group.
5. Let them play your story.
6. Score their performance using your rules. Do this manually, reading the log files yourself. In this process, it is permissible to correct the rules if necessary. When 20 persons play, exceptions will occur that are not yet covered by your rules.
7. Using your statistical software, check if there are any tendencies towards correlation among your rules. This means that if two rules are going to measure the same skill, then the list of scores achieved by the users for these two rules should be related statistically. You should not expect significant correlations, but you can expect non-significant positive correlations. Just as an example, for two rules, the significance can be around 0.100. Sometimes, there is no significance, and sometimes there is a high one, but in all cases, the correlation coefficient should not be negative. What if the correlation is negative or even significantly negative? Try modifying the rule. If the situation persists, you must delete the behavioural marker entirely as it

does not contribute to assessing the same skill as the others do. However, you might find two or more blocks within your set of rules: within the block, they have some correlation; across the blocks, there is no or even a negative relationship. If this is the latter case and you are measuring two skills in the same story, that is fine. However, this is not ideal because you are measuring both skills with fewer behavioural markers than recommended. This will result in fewer chances to validate the assessment (Step 10).

8. Finally, document the successful rules. It is a reasonable expectation to arrive at around eight rules (each corresponding to one behavioural marker). If there are blocks as mentioned above, document them as tests for separate skills.

Step 10: Validation of assessment

Time: 1 month of full-time work distributed over 2 months of time. Requisites: 50 people of the target group, 1 assessment centre role-player, 2 assessment centre assessors, 1 audio recorder that records in good quality when placed at the middle of a table, and statistical software to calculate correlations.

This step is to validate that your story truly measures what it should. In other words, if your story results in the judgment that a person is good or bad at a skill, then this should correspond to the real-world ability of this person with regard to the same skill.

This step is not described here: conducting assessment centres is a profession of its own (assessment psychologist) and cannot be learned simply by reading a description.

8 Literature

Ahmed, Y., Payne, T. & Whiddett, S. (1997). A process for assessment exercise design: A model of best practice. International Journal of Selection and Assessment, 5: 62-68.

Archer, D. & Akert, R. M. (1977). Words and everything else: Verbal and nonverbal cues in social interpretation. Journal of Personality and Social Psychology, 35: 443-449.

Austin, J. L. (1962) How to Do Things With Words. Cambridge, MA: Harvard University Press.

Bartram, D. (2004). Assessment in Organisations. Applied Psychology, 53: 237–259.

Baddeley, A. D. & Weiskrantz, L. (eds.) (1993). Attention, Selection, Awareness and Control. Oxford: Clarendon Press.

Blake, R. & Mouton, J. (1964). The Managerial Grid: The Key to Leadership Excellence. Houston: Gulf Publishing Co.

Blake, R. & Mouton, J. (1985). The Managerial Grid III: The Key to Leadership Excellence. Houston: Gulf Publishing Co.

Bobko, P., Roth, P. L. & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors and job performance. Personnel Psychology, 52: 561–589.

Brändli, M. (2008). Erste empirische Datenkollektion des PM-Game und dessen Graphical User Interface. Unterlagen zur Studie. Vertiefungsarbeit am Institut für Informatik, Universität Zürich.

Breitner, M. (ed.) (2006). E-Learning Geschäftsmodelle und Einsatzkonzepte, Wiesbaden: Gabler.

Brummel, B. J., Rupp, D. E. & Spain, S. M. (2009). Constructing parallel simulation exercises for assessment centres and other forms of behavioural assessment. Personnel Psychology, 62: 137–170.

Bray, D. W., R. J. Campbell, & D. L. Grant. (1974). Formative Years in Business: A Long-Term AT&T Study of Managerial Lives. New York: Wiley & Sons.

Bycio, P., Alvares, K. M. & Hahn, J. (1987). Situational specifity in assessment center ratings: A confirmatory factor analysis. Journal of Applied Psychology, 72: 463–474.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, Psychological Bulletin, 56: 81-105.

Chan, D. & Schmitt, N. (2005). Situational judgment tests. In Evers, A., Anderson, N. & Voskuijl, O. (eds.), The Blackwell Handbook of Personnel Selection. Malden, MA: Blackwell, 219-42.

Christel, M. (1994). The role of visual fidelity in computer-based instruction, Human-Computer Interaction, 9: 183-223.

Conway, A. R., Kane, M. J., & Engle, RW (2003). Working memory capacity and its relation to general intelligence. Trends in Cognitive Sciences, 7 (12): 547–52.

Corbett, A. T. (2001). Cognitive computer tutors: Solving the two-sigma problem. User Modeling: Proceedings of the Eighth International Conference, UM 2001, 137-147.

Costa, P. T., Jr. & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. Odessa, FL: Psychological Assessment Resources.

Croux C., Dehon, C. (2010). Influence Functions of the Spearman and Kendall Correlation Measures. CentER Discussion Paper Series No. 2010-40.

Donovan, M.A. (1998) Interactive video assessment of conflict resolution skills. Personnel Psychology, 51 (1): 1-24.

Davis, B. (2009). On Designing Controlled Natural Language for Semantic Annotation. Appeared in: Fuchs, N. Controlled Natural Language: Workshop on Controlled Natural Language, CNL 2009, Marettimo Island, Italy, June 8-10, 2009. LNCS/LNAI 5972, Springer, 2010.

Dede, C. (1995). Assessment of emerging educational technologies that might assist and enhance school-to-work transitions. Washington, D.C.: National Technical Information Service. Report to the United States Congress.

Denning, P. J. (1997). A New Social Contract for Research. Communications of the ACM, 40 (2): 132-134.

Diercks, J., Eingel, S., Jägeler, T., & Weber, A. (2003). Vorteile und Nutzenpotenziale kombinierter-Recruiting- und Marketinganwendungen. Ein Praxisbeispiel für Recrutainment: CYQUEST 'Die Karrierejagd durchs Netz'. In U. Konradt & W. Sarges (Eds.), E-Recruitment und E-Assessment. Göttingen: Hogrefe: 127-144.

Drasgow, F., & Olson-Buchanan, J. B. (1999). Innovations in computerized assessment. Mahwah, NJ: Lawrence Erlbaum.

Drasgow, F., Olson, J. B., Keenan, P. A., Moberg, P., & Mead, A. D. (1993). Computerized assessment. Research in Personnel and Human Resources Management, 11: 163-206.

Ebert, J. (2008). Assessment Center Method to Evaluate Practice-Related University Courses - Challenges & Chances. Details of the presentation at ENAC, August 27-29, 2008. Potsdam, Germany.

Emerson, E. N.; Crowley, S. L. & Merrell, K. W. (1994). Convergent validity of the School Social Behavior Scales with the Child Behavior Checklist and Teacher's Report Form. Journal of Psychoeducational Assessment, 12 (4): 372-380.

Escudier, M., Newton, T., Cox, M., Reynolds, P. & Odell, E. (2011). University students' attainment and perceptions of computer delivered assessment; a comparison between computer-based and traditional tests in a 'high-stakes' examination. Journal of Computer Assisted Learning, 27.

Evers, A, et al. (eds.), (2005) Handbook of selection. Malden MA: Blackwell Publishing: 419-439.

Farrell, D., M. Laboissière, J. Rosenfeld, S. Stürze, & Umezawa, F. (2005). The emerging global labor market: Part II—the supply of offshore talent. San Francisco, CA: McKinsey Global Institute.

Feldman, R. S., Philippot, P., & Custrini, R. J. (1991). Social competence and nonverbal behavior. In R. S. Feldman, & Rim., B. (eds.), Fundamentals of nonverbal behavior. Cambridge MA: Cambridge University Press: 329-350.

Ferris, G. R., Perrewé, P. L. & Douglas, C. (2002). Social effectiveness in organizations: Construct validity and research directions. Journal of Leadership and Organizational Studies, 9: 49-63.

Ferris, G. R., Witt, L. A., & Hochwarter, W. A. (2001). Interaction of social skill and general mental ability on job performance and salary. Journal of Applied Psychology, 86: 1075-1082.

Fleenor, J.W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. Journal of Business and Psychology, 10: 319-333.

Fletcher, C., Baldry, C. & Cunningham-Snell, N. (1998). The Psychometric Properties of 360 Degree Feedback: An Empirical Study and a Cautionary Tale. International Journal of Selection and Assessment, 6: 19–34.

Fletcher, G., Flin, R., McGeorge, P. , Glavin R., Maran N., Patey R. (2004). Rating non-technical skills: developing a behavioural marker system for use in anaesthesia. Cogn Tech Work, 6: 165–171.

Florescu, D., Kossmann, D. (2008). Rethinking the Cost and Performance of Database Systems. SIGMOD Record, 38 (1): 43-48.

Ford, M. E. (1982). Social cognition and social competence in adolescence. Developmental Psychology, 18: 323-340.

Funder, D. C. & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. Journal of Personality and Social Psychology, 64: 479-490.

Funke, L. (1998) Computer-based testing and training with scenarios from complex problem solving research: Advantages and disadvantages. International Journal of Selection and Assessment, 6: 90-96.

Funke, U. & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. International Journal of Selection Assessment, 6: 115-23.

Gaugler, B.B. & Thornton, G.C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. Journal of Applied Psychology, 74: 611-618.

Gereffi, G., Wadhwa, V., Rissing, B., & Ong, R. (2008). Getting the Numbers Right: International Engineering Education in the United States, China , and India. Journal of Engineering Education, 97: 13-25.

German Federal Ministry of Education and Research (1999). New Approaches to the Education and Qualification of Engineers: Challenges and Solutions from a Transatlantic Perspective.

Gerpott, T. J. (1990). Erfolgswirkungen von Personalauswahlverfahren. Zur Bestimmung des ökonomischen Nutzens von Auswahlverfahren als Instrument des Personalcontrolling. Zeitschrift für Organisationspsychologie, 1: 37-44.

Gollwitzer, P. M. & Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. Journal of Personality and Social Psychology, 73, 186-199.

Greguras, G. J., Robie, J & Born, M. P. (2001). Applying the social relations model to self and peer evaluations. Journal of Management Development, 20: 508-525.

Greyling, L., Visser, D. & Fourie, L. (2007). Construct validity of competency dimensions in a team leader assessment centre. SA Journal of Industrial Psychology, 29 (2): 10-19.

Hargie, O. (2006). The handbook of communication skills. 3rd ed. London: Routledge.

Harris, M. M. & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41: 43-62.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (eds.) Human Mental Workload. Amsterdam: North Holland Press.

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting. Santa Monica: HFES: 904-908.

Hartmann, D.P., Roper, B.L. & Bradford, D.C. (1979). Some relationships between behavioral and traditional assessment. Journal of Behavioral Assessment, 1: 3–21.

Hasler, B. S. (2009). Virtual assessment center. A media comparison study on the diagnostic value of online role-play for social competence assessment. Doctoral dissertation at the University of Zurich. Marburg, Germany: Tectum.

Hennessy, J., Mabey, B. & Warr, P. (1998), Assessment Centre Observation Procedures: An Experimental Comparison of Traditional, Checklist and Coding Methods. International Journal of Selection and Assessment, 6: 222–231.

Hertel, G., Konradt, U., & Orlikowski, B. (2003). Ziele und Strategien von E-Assessment aus Sicht der psychologischen Personalauswahl. In U. Konradt & W. Sarges (Hrsg.), E-Recruitment und E-Assessment. Göttingen: Hogrefe: 37-53.

Hesketh, A. (2000). Recruiting an elite? Employers' perceptions of graduate education and training. Journal of Education and Work 13: 245–271.

Hess, M., Mahlow, C. (2007). Sentence Completion Tests in a Virtual Laboratory. German e-Science Conference 2007, May, Baden-Baden, Germany: 1-10.

Hertel, G., Konradt, U. & Orlikowski, B. (2003). Ziele und Strategien von E-Assessment aus Sicht der psychologischen Personalauswahl. In U. Konradt & W. Sarges (Hrsg.), E-Recruitment und E-Assessment. Göttingen: Hogrefe: 37-53.

Hevner, A.R., S.T. March, J Park, & S. Ram (2004). Design science in information systems research. MIS Quarterly, 28: 75-105.

Höft, S. & Schuler, H. (2001). The conceptual basis of assessment centre ratings. International Journal of Selection and Assessment, 9: 114-123.

Hogan, J. & Zenke, L.L. (1986). Dollar-value utility of alternative procedures for selecting school principals. Educational and Psychological Measurement, 46: 935-945.

Hoffmann, C.C & Thornton, G.C. (1997). Examining selection utility where competing predictors differ in adverse impact. Personnel Psychology, 50: 455-470.

Holaday M, Smith DA, Sherry A. (2000). Sentence completion tests: a review of the literature and results of a survey of members of the Society for Personality Assessment. J Pers Assess. 74 (3): 371-83.

Hong, J.-C., Cheng, C.-L., Hwang, M.-Y., Lee, C.-K. & Chang, H.-Y. (2009). Assessing the educational values of digital games. Journal of Computer Assisted Learning, 25: 423–437.

Irvine, S., & Kyllonen, P. (eds.) (2002). Item generation for test development. Hillsdale, NJ: Lawrence Erlbaum.

Ito, T. (2009). Das Evolutionäre Lernspiel-Konzept: Eine Kombination aus Game-based Learning und Web 2.0. Dissertation at the Faculty of Economical Sciences, Zurich University.

Jackson, D.J.R., Stillman, J.A. & Atkins, S.G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. Human Performance, 18: 213–241.

Jackson, D. J. R., Stillman, J. A. & Englert, P. (2010). Task-Based Assessment Centers: Empirical support for a systems model. International Journal of Selection and Assessment, 18: 141–154.

Kaufmann, E. (2009). Talking to the Semantic Web: natural language query interfaces for casual end-users. Dissertation at the University of Zurich.

Karkoschka, U. (1998). Validität eignungsdiagnostischer Verfahren zur Messung sozialer Kompetenz. Frankfurt am Main: Peter Lang Verlag.

Keenan, A. (1977). Some relationships between interviewers' personal feelings about candidates and their general evaluation of them. Journal of Occupational Psychology, 50: 275-283.

Kelbetz, G. & Schuler, H. (2002). Verbessert Vorerfahrung die Leistung im Assessment Center? Zeitschrift für Personalpsychologie, 1: 4–18.

Keller, M. (2008). Konkurrenzanalyse für Projektmanagement Games; diploma thesis, University of Zurich.

Klampfer B., Flin R., Helmreich R.L., Hausler R., Sexton B., Fletcher G. et al (2001). Enhancing performance in high risk environments: recommendations for the use of behavioural markers. Ladenburg: Daimler-Benz Stiftung.

Klein, C., DeRouin, R. E., & Salas, E. (2006). Uncovering workplace interpersonal skills: A review, framework, and research agenda. G. P. Hodkinson & J. K. Ford (eds.), International review of industrial and organizational psychology. West Sussex, UK: Wiley: 79-126.

Kleinmann, M. (1997). Assessment Center. Stand der Forschung - Konsequenzen für die Praxis. Göttingen: Verlag für Angewandte Psychologie.

Kleinmann, M. (2003). Assessment-Center. Göttingen: Hogrefe.

Kleinmann, M. & Strauss, B. (1998). Validity and application of computer-simulated scenarios in personnel assessment. International Journal of Selection and Assessment, 6: 97-106.

Konradt, U. & Sarges, W. (2003). E-Recruitment und E-Assessment. Göttingen:Hogrefe.

Kudisch, J. D., Ladd, R. T. & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. Journal of Social Behavior and Personality, 12: 129–144.

Kuhl, J. (2001). Motivation und Persönlichkeit. Interaktionen psychischer Systeme. Göttingen: Hogrefe.

Lance, C. E. (2008) Why assessment centers do not work the way they are supposed to. Industrial and Organizational Psychology: Perspectives on Science and Practice, 1: 84–97.

Laumer, S., von Stetten A. & Eckhardt, A. (2009). E-Assessment. Wirtschaftsinformatik, 51 (3): 306-308.

Laver, J. (1975). Communicative Functions of Phatic Communion, in: Kendon, A. / Harris, R. / Key, M. (eds.), The Organisation of Behaviour in Face-to-Face Interaction, The Hague: Mouton: 215–238.

Lee, F. J. & Anderson, J. R. (1997). Learning to Act: Acquisition and optimization of procedural skill. In Proceedings of the 19th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum: 418-423.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R). Manual (2. erweiterte und überarbeitete Aufl.). Göttingen: Hogrefe.

Lievens, F. & Van Keer, E. (2001). The construct validity of a Belgian assessment centre: A comparison of different models. Journal of Occupational and Organizational Psychology, 74: 373–378.

Luecke, R. & Hall, B.J. (2006). Performance management: Measure and improve the effectiveness of your employees. Boston, MA: Harvard Business School Press.

Mahlow, C. & Hess, M (2004). Sentence completion tests for training and assessment in a computational linguistics curriculum. In: COLING-2004 Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning, Geneva, Switzerland, August 2004: 61-70.

Malinowski, B. (1923). The problem of meaning in primitive languages. In: Ogden, C. & Richards, I., The Meaning of Meaning. London: Routledge.

Mason, G. (2004). Enterprise product strategies and employer demand for skills in Britain. *SKOPE Working Paper* 50. London: National Institute for Social and Economic Research.

Mateas, M. & Stern, A. (2003). Facade: An Experiment in Building a Fully-Realized Interactive Drama. In Game Developer's Conference: Game Design Track, San Jose, California.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L & Grubb, W.B. (2007). Situational judgment tests, response instructions and validity: a meta-analysis. In Burke, M.J. (Hsg.). Personnel psychology. Blackwell Publishing: 63-84.

McKee, R.; Carlson, B. (1999). The Power to Change. Austin, Texas: Grid International Inc.

Newell, A., & Simon, H. (1976) ‘Computer Science as Empirical Inquiry: Symbols and Search,’ Communications of the ACM, 19 (3): 113-126.

Mehrabian, A. & Wiener, M. (1967). Decoding of inconsistent communications. Journal of Personality and Social Psychology, 6: 109-114.

MeritTrac (2007): Engineering Graduate Pool of India. Bangalore: MeritTrac Inc.

Merrell, K. W. & Caldarella, P. (1999). Social-behavioral assessment of at-risk early adolescent students: Psychometric characterstics and validity of a parent report form of the School Social Behavior Scales. Journal of Psychoeducational Assesment, 17 (1): 36-49.

Mischel, W. & Shoda, Y. (1995). A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure. Psychological Review, 102 (2): 246-268.

Moses, J. L. (1973). The development of an assessment center for the early identification of supervisory potential. Personnel Psychology, 26: 569–580.

Moss, F.A., & Hunt, T. (1927). Are you socially intelligent? Scientific American, 137: 108-110.

Motowidlo, S. J., Dunnette, M. D. & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. Journal of Applied Psychology, 75: 640–647

Myers-Briggs, I., & Myers, P. (1980, 1995). Gifts Differing: Understanding Personality Type. Mountain View, CA: Davies-Black Publishing.

Ohlsson, S. (1994). Constraint-based Student Modeling. Student Modeling: the Key to Individualized Knowledge--based Instruction. Berlin: Springer-Verlag: 167-189.

O'Neil, H. F., Allred, K., & Dennis, R. A. (1994). Assessment Issues in the Validation of a Computer Simulation of Negotiation Skills. CSE Technical Report 374, April 1994. University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

O'Neil, H. F., Allred, K., & Dennis, R. A. (1997). Validation of a computer simulation for assessment of interpersonal skills. In H. F. O'Neil (ed.), Workforce readiness: Competencies and assessment. Mahwah, NJ: Lawrence Erlbaum: 229-254.

Oppenheimer, L. (1989). The nature of social action: Social competence versus social conformism. In Schneider, G. Atilli, J., Nadel, Weissberg, R. (eds.), (1989). Social competence in developmental perspective. Dordrecht, The Netherlands: Kluwer International Publishers.

Paschall, M. J., Fishbein, D. H., Hubal, R. C. & Eldreth, D. (2005). Psychometric properties of virtual reality vignette performance measures: a novel approach for assessing adolescents' social competency skills Health Educ. Res. 20 (1): 61-70.

Pass, F., Tuovinen J. E., Tabbers, H., Van Gerven, Pascal, W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory, Educational Psychologist, 38 (1): 63-71.

Probst, P. (1975). Eine empirische Untersuchung zum Konstrukt der Sozialen Intelligenz. Diagnostica, 21: 24–27.

Raiter, M. & Warner, D. E. (2005). Social Context in Massively-Multiplayer Online Games (MMOGs): Ethical Questions in Shared Space. International Review of Information Ethics, 4 (12): 46-52.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV. Berkeley, California: University of California Press: 321–334.

Reilly, R., Henry, S. & Smither, J. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. Personnel Psychology 43: 71-84.

Resta, P. & Laferrière, T. (2007). Technology in Support of Collaborative Learning. Educational Psychology Review, 19: 65–83.

Riggio, R. E. (1986) Assessment of basic social skills. Journal of Personality and Social Psychology, 51: 649-660.

Riggio, R. E., Riggio, H. R., Salinas, C., & Cole, E. J. (2003). The role of social and emotional communication skills in leader emergence and effectiveness. Group Dynamics: Theory, Research, and Practice, 7: 83-103.

Rose-Krasnor, L. (1997). The nature of social competence: A theoretical review. Social Development, 6: 111-135.

Searle, J. (1969). Speech Acts: An Essay in the Philosophy of Language, Cambridge, England: Cambridge University Press.

Segrin, C. (1998). The impact of assessment procedures on the relationship between paper and pencil and behavioral indicators of social skill. Journal of Nonverbal Behavior, 22: 229-251.

Spitzberg, B. H., & Cupach, W. R. (1989). Handbook of interpersonal competence research. New York: Springer.

Sternberg, R. J. & Smith, L. (1985). Social intelligence and decoding skills in non-verbal communication. Social Cognition, 3: 168–192.

StepStone DeutschlandAG (2004). Studie: Aktuelle Trends auf dem Bewerbermarkt. Düsseldorf: StepStone Deutschland AG.

Stevens, S. (1989), Intelligent interactive video simulation of a code inspection. Communications of the ACM, 32 (7): 832-843.

Stoyan, R. (2008): 'PM for all': Intensive small group teaching in project management, for many students at low cost. International Journal of Project Management, 26: 297–303.

Strauss, B. & Kleinmann, M. (1995). Computersimulierte Szenarien in der Personalarbeit. Göttingen: Verlag für Angewandte Psychologie.

Szilas, N. (2007). A Computational Model of an Intelligent Narrator for Interactive Narratives. Applied Artificial Intelligence, 21 (8): 753-801.

Tennant, H. et al. (1983). Menu-Based Natural Language Understanding. 21st Meeting of the Assoc. for Computational Linguistics, Assoc. for Computational Linguistics, MIT: 151–158.

Tennant, H. R. (1980). Evaluation of natural language processors. Dissertation at the Department of Computer Science, University of Illinois.

Thompson, C., Pazandak, P. & Tennant, H. (2005). Talk to your semantic Web. Internet Computing, IEEE, 9: 75-78.

Thorndike, R. L. & Stein, S. (1937). An evaluation of the attempts to measure social Intelligence. Psychological Bulletin, 34: 275–285.

Thornton, G. C. (1992). Assessment Centres in Human Resource Management. Reading, MA: Addison-Wesley.

Tsichritzis, D. (1998). The Dynamics of Innovation. In Beyond Calculation: The Next Fifty Years of Computing, P. J. Denning & R. M. Metcalfe (eds.), NewYork: Copernicus Books: 259-265.

Zedeck, S. (1986). A process analysis of the assessment center method. In B.M. Staw & L.L. Cummings (eds.), Research in Organizational Behavior, 8: 259-296. Greenwich, CT: JAI Press.

Wells, D. (2011). Simplicity is key. Downloaded from http://www.extremeprogramming.org/rules/simple.html on the 1th of December 2011.

Wiener, M. (2006). Critical success factors of offshore software development projects, Wiesbaden: Deutscher Universitäts-Verlag.

9 Curriculum Vitae

Robert Stoyan
robert.stoyan@gmail.com

Research and Industry Career

TECFA, FPSE, Université de Genève
Since 2011

Leader of the research project CultuMento. The aim of the project is to investigate personal coaching in intercultural skills with the PM Game.

Institute of Psychology, Osnabrück University, Germany
07.2009 – 06.2010

Guest Researcher at the Chair of Personal Psychology, Prof. Julius Kuhl

Institute for Informatics, University of Zurich, Switzerland

2004-2010

Research, teaching and project management at the Chair of Educational Engineering, Prof. Helmut Schauer

- 2007-2009 Leader of the research project 'PM Game' on e-assessment of communicative abilities (the dissertation project presented in this thesis)
- 2004-2008 Leader of the research project 'PM for all' on interactive face-to-face teaching of team leadership and project management abilities

Industry career 1995-2003 with employments in Project Management and Software Development at:
GFT AG, Frankfurt, Germany
Entory AG, Karlsbad, Frankfurt, Germany
imbus GmbH, Möhrendorf, Germany